This is to certify that the
dissertation entitled

THE ASYMPTOTIC DISTRIBUTION OF AN IRT MEASURE FOR ITEM FIT BASED ON PSEUDOCOUNTS

presented by

Deping Li

has been accepted towards fulfillment
of the requirements for the

PH.D. degree in Education

Mark D. Reckase
Major Professor's Signature

August 2, 2005
Date

MSU is an Affirmative Action/Equal Opportunity Institution

THE ASYMPTOTIC DISTRIBUTION OF AN IRT MEASURE FOR ITEM FIT BASED ON PSEUDOCOUNTS

By

Deping Li

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counselling, Educational Psychology and Special Education

2005

ABSTRACT

THE ASYMPTOTIC DISTRIBUTION OF AN IRT MEASURE FOR ITEM FIT BASED ON PSEUDOCOUNTS

By

Deping Li

Item fit measure Q^*_{DM} is formed based on the posterior distribution (or pseudocounts) of proficiency instead of the proficiency estimates. The reference distribution of Q^*_{DM} is not χ^2 but a quadratic function of normal variates. A consistent estimator of the covariance matrix of pseudocounts is found for the approximation of the true asymptotic distribution of Q^*_{DM}. The data-based estimate of the covariance matrix of pseudocounts depicts the interrelations among pseudocounts and show reasonably good agreement with the true covariance matrix among pseudocounts for sample size as large as 1000. Results from simulation studies show that the method based on pseudocounts has adequate power for detecting item misfit and low type I error rates. The method is robust over the underlying ability distribution and number of quadrature points. Real data applications suggest that the method provide more helpful information on assessing model-data fit even when sample size is large compared to χ^2 test.

ACKNOWLEDGEMENTS

I am indebted to many people for criticism, suggestions, reviews, and constructive conversations. I wish to express my sincere thanks to the committee: Dr. Mark Reckase (chair), Dr. Kimberly Maier, Dr. Lijian Yang, and Dr. John Donoghue. Each contributed tremendously to the work by sharing their extensive professional knowledge and ideas.

I am especially grateful to Dr. Donoghue and Dr. Catherine McClellan for the continued advise, criticism, and wisdom, beginning from the summer research experience through the completion of this work. I would like to thank Educational Testing Service for their financial support, through both summer intern research and the fellowship offered for this research.

I would also like to thank the Center for Educational Performance and Information and Dr. Oren Christmas, whose assistance enabled the completion of my doctoral study. Thanks also are due Hongwen Guo for her insightful critiques and helpful comments.

The encouragement by my wife, Yanlin Jiang, and her support in all aspects did much to reduce the burden of the work involved.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to IRT Measures of Item Fit

1.1 Item Fit in General Context of Assessing the Fit of the IRT models

Item response theory (IRT) is becoming an important tool for educational and psychological tests, one of the most important tools for both test design and test data analysis. IRT provides a philosophical framework for test design and many other applications (e.g., differential item functioning, test equating, computer adaptive testing, etc.). The advantages of IRT may not be fully realized if the test data do not adequately fit the item response models. Assessing model-data fit is fundamental in psychometrics and has always been an issue of enormous interests. The model-data fit issue should be a primary concern when applying IRT models to test data. However, there is no unanimous consensus upon the diagnostic tools for model-data fit.

There are other aspects of model-data fit (e.g., person fit analysis and analysis of other type of misfit including violation of local independence and unidimensionality by Hambleton & Swanminathan (pp. 151-195), 1985; Embreston & Reise (pp. 238-

246), 2000; Glas & Meijer, 2003, Hoijtink 2001; and Sinharay and Johnson 2003), but this research is limited to item fit only. In IRT, there is no need to fit a set of data with the same model for all items because a test can be a combination of different types of items (e.g., dichotomous, polytomous, or constructed response items). Even if items with the same type of responses are available, they may be represented by different mathematical models, and separate IRT models may be used for adequate fit. Therefore, attention should be paid to the fit of IRT model on an item-by-item basis.

Item fit analysis should also play an important role in decisions about the retention of items in the assessment pool. Poorly fitting items undermine the validity of decisions based on measurement results. In this chapter, various measures of item fit and the corresponding statistical approaches for testing goodness-of-fit at the item level will be reviewed.

Generally speaking, there are two basic approaches to assessing item fit—graphical (or heuristic) and statistical test procedures. Graphical procedures are intuitive but more subjective in deciding the adequacy of model-data fit. Statistical tests of goodness of fit (e.g., χ^2 or likelihood ratio test) are probably the most widely used in current operational research.

In graphical procedures, the adequacy of item fit is typically evaluated on the basis of a comparison between an empirical item response function and a hypothetical item response function. The empirical function is obtained from the sample of test data. Detailed descriptions of graphical procedures can be found in most IRT literature

dwelling on model-data fit (e.g., Hambleton & Swaninathan, 1985; p234, Embreston & Reise, 2000). The plots of the empirical and hypothetical item response functions can reveal areas along the proficiency continuum where there are discrepancies between these two functions. The discrepancies indicate the degree of item misfit.

1.2 Item Fit Analysis Based on Ability Estimates

Much research on analysis of item fit has been conducted via significance tests. This section reviews Wright and Panchapakesan's (1969) χ^2 test, Bock's (1972) χ^2 test, likelihood ratio test, and standardized residuals test.

The procedure advocated by Wright and Panchapakesan (1969) is a commonly used statistical test. The procedure defines a standardized variable $y_{ij} = \left(\frac{f_{ij}-Ef_{ij}}{\sqrt{Var(f_{ij})}}\right)$, where f_{ij} represents the frequency of examines at the ith ability level answering the jth item correctly. Then the measure of item fit $\chi^2 = \sum_{g=1}^{G} y_{ij}^2$. Wright and his colleagues assume this measure to have a chi-square distribution.

The Bock (1972) chi-square index is defined as

$$\chi^2_{Bock} = \sum_{g=1}^{G} \frac{N_g(O_{ig} - E_{ig})^2}{E_{ig}(1 - E_{ig})},$$

where O_{ig} is observed proportion-correct on item i for interval group g, E_{ig} is the expected proportion correct based on the hypothetical item response function at the within interval median proficiency level estimate, and N_g is the number of examinees with ability estimates falling within proficiency interval g that comes from the classification of the proficiency estimates. This index is assumed to distribute asymp-

totically as a χ^2 variable with degree of freedom equal to $G - m$, where m represents the number of item parameters to be estimated. High value of the item fit index indicate that the data may not have a reasonable with fit the hypothetical *model* on the item.

The Wright and Mead (1977) statistic is based on number-correct grouping approach for Rasch model. The statistic is given by

$$\chi^2 = \sum_{g=1}^{G} \frac{N_g(O_{ig} - E_{ig})^2}{E_{ig}(1 - E_{ig}) - S_{pj}^2},$$

where $S_{pj}^2 = \frac{1}{N}\sum_{k\in j}^{N_j}(P_i(\theta_k) - E_{ij})^2$, $P_i(\theta_k)$ is the proportion correctly answering item i in score group k. The degrees of freedom are G, the number of intervals for the proficiency estimates, minus the number of parameters estimated.

Yen's (1981) Q_1 statistic uses the mean proficiency within each proficiency category to obtain the predicted item response function. Furthermore, Yen fixes 10 categories of proficiency in calculating the Q_1 index, which is assumed approximately distributed as χ^2 with the number of categories minus the number of parameters as the degree of freedom.

The likelihood ratio G^2 is implemented in the BILOG-3 (Mislevy and Bock, 1990) and BILOG-MG (Zimowski, Muraki, Mislevy, and Bock, 1996). G^2 is computed by comparing the observed frequencies with those predicted from the hypothetical model.

$$G_{BILOG}^2 = 2\sum_{i=1}^{I}\left(R_i log\frac{R_i}{N_i(P(\theta_m))} + (N_i - R_i)log\frac{R_i}{N_i(1 - P(\theta_m))}\right).$$

This test of item fit was designed from a long test (e.g., more than 20 items). In

this test, EAP estimate of proficiency for each examinee is computed based on the item parameter estimates, then is assigned to proficiency intervals. The summation is performed over G ability scale θ groups, R_i is the proportion correct within group i, and N is the number of examinees in group i. This G^2 is also assumed to distributed as χ^2 with the degrees of freedom equal to the number of proficiency groups.

Standardized residuals are used to assess the item fit in the Rasch model context (e.g., Masters & Wrights, 1996). In this procedure, the expected response EX_{si} for a particular person s responding to item i is described by $EX_{si} = \sum_{k=1}^{K-1} kP_i(\theta_s)$. The variance of X_{si} can be calculated by $Var(X_{si}) = \sum_{k=0}^{K-1}(k - EX_{si})^2 P_i(\theta_s)$. Let Z_{si} denote the standardized residual, then $Z_{si} = \frac{X_{si} - EX_{si}}{\sqrt{Var(X_{si})}}$. A mean square fit statistic, i.e., $\sum_{i=1}^{n} \frac{Z_{si}^2}{n}$, can then be computed as an item fit measure. The summation is performed over the n items in the test.

The above measures of item fit and corresponding statistical tests are open to criticisms. The most common criticism is that these item fit measures and the corresponding significance tests often require parameter estimation (i.e., item and ability estimates) and are often viewed as inconclusive evidence of adequate fit. The most commonly used measures of item fit (e.g., Bock, 1972; Yen, 1981) use model-based estimates (e.g., maximum likelihood estimate (MLE), or expectation a posterior (EAP) of the latent proficiency of examinees. In computing these fit measures, the proficiency estimates are generally treated as point estimates containing no error—an obviously false assumption. That is, even if there is perfect fit of the model to the data, the proficiency estimate for an individual is hardly ever equal to the true value

due to estimation errors. This problem is especially pronounced for short tests where proficiency estimates have larger error. In addition, the proficiency estimates are then grouped into intervals that serve as the basis of a contingency table measure of fit. Due to the uncertainties in the proficiency estimation, the proficiency estimates are subject to errors of classification, thus making the use of the chi-square reference distribution questionable. Several studies (e.g., Reise, 1980; Rogers and Hattie, 1987; Mckinley and Mills, 1985) have indicated that the sampling distributions of these measures are not χ^2 distributed. Moreover in some contexts, researchers point out that the χ^2 statistic for a single item is insensitive to certain type of misfit (e.g., Vander Wollenberg, 1982; Drasgow et al 1995).

1.3 Item Fit Analysis Based on Raw Scores

Because of the shortcomings of measures based on point estimates of ability, alternative measures have been developed. In the past 10 years, two main approaches have been put forth. The first approach was suggested by Orlando and Thissen (2000, 2003). Their approaches compute IRT-based expected values for each level of total score on the test, raw score or number correct score. They then use the observed frequencies for the total scores, and compute a fit measure (likelihood ratio G^2 or Pearson χ^2). The item fit statistics for item i suggested by Orlando and Thissen

(2000) are of the form

$$S - \chi_i^2 = \sum_{k=1}^{I-1} N_k \frac{(p_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}$$

and

$$S - G_i^2 = 2 \sum_{k=1}^{I-1} N_k [p_{ik} log(\frac{p_{ik}}{E_{ik}}) + (1 - p_{ik}) log(\frac{1 - p_{ik}}{1 - E_{ik}})],$$

with k standing for raw score category as $k = 0, 1, 2, \cdots, I$, N_k for the number of examinees on score k, p_{ik} and E_{ik} respectively representing the observed and expected correct scores for item i in raw score group k. Orlando and Thissen then compare the statistic to a chi-square distribution (the two statistics are assumed to have asymptotic $\chi^2(I-4)$ distributions under the null hypothesis that the fitted model is true). Unfortunately, their statistic is not distributed exactly as chi-square when item parameters are estimated from MMLE (Donoghue, McClellan, and Oranje, 2004; Sinharay, 2005). However, the departure from χ^2 appears to be relatively small, a result supported by several simulation studies (e.g., Orlando and Thissen, 2000; Stone and Zhang, 2002); the departure of the distribution of $S - \chi^2$ and $S - G^2$ from the referred $\chi^2(I - 4)$ distribution may be severe for a short test.

Glas and Suarez-Falcon (2003) suggest an item fit statistic based on the lagrange multiplier test (or equivalent efficient score test) and uses number correct score on examinee groups. For item i, the statistic is used to test the null hypothesis H_o (e.g.,

the 3PL model is correct) versus the alternative hypothesis, in which the model is defined as

$$p(u_i|\theta, a_i, b_i, c_i, \beta_{is}, s) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta - b_i - \beta_{is})}},$$

where s indicates the raw score group an examinee belongs to, a_i, b_i, c_i describe the parameters for item i, and β_{is} adjusts the item difficulty b_i from the score group s. The test statistics, which is defined as $\mathbf{h_i'}\Sigma\mathbf{h_i}$, has an asymptotic $\chi^2(S_i - 1)$ distribution. In computing the test statistic, $\mathbf{h_i}$ is a vector of differences between the observed proportion correct and its posterior expectation for a raw score group computed based on MMLE, and Σ_i is the estimated matrix of $\mathbf{h_i}$. Even though this test statistic appears to have a strong theoretic basis, Glas and Suarez-Falcon (2003, p.97) found that overall characteristics of their test statistic is worse then that of $S - \chi^2$ and $G - \chi^2$. Researchers (e.g., Sinharay, 2005) points out that assessing item fit using number correct score on examinee groups is not entirely satisfactory and there is a substantial scope of further research in this area.

Recently, Sinharay (2005) from a Bayesian perspective suggested uses of the χ^2-type and G^2-type test statistics of Orlando and Thissen (2000) as a summary measure of discrepancy, but computed the posterior predictive distributions as the reference distributions. The resulting Bayesian p-values provide probability statements about the fit of the data with the model on the items. This method also has strong theoretic basis. However, the posterior predictive model checking methods are heavily dependent on the resampling methods and are using the MCMC algorithm and hence

are computationally intensive.

1.4 Item Fit Analysis Based on Pseudocounts

The second approach of a fit measure called Q^*_{DM}, is proposed by Donoghue and McClellan (e.g., 2004, 2003b, 2003a, 2001b, 2001a, 1999). In this approach, the asymptotic distribution of an alternative IRT measure of item fit, referred to as Q_{DM}, is derived and well justified as asymptotically quadratic form of normal variables. Q^*_{DM} is based on pseudocounts as opposed to counting the number of examinees falling within a proficiency interval on the basis of proficiency estimates. It is a natural by-product of the MML-EM estimation (Bock and Lieberman, 1970; Bock and Aitkin, 1981) used by most IRT calibration programs. This measure has generated much study (e.g., Stone, 2000; Stone, Ankerman, Lane, and Liu, 1993; Stone and Hansen, 2000; Stone, Mislevy and Mazzeo, 1994; Stone and Zhang, 2002 Donoghue and Isham 1998; and Donoghue and Hombo, 1999, 2001ab, 2003ab; Hombo and Donoghue, 1999, 2000, 2001; Hombo, Donoghue and Oranje, 2003). Simulation studies (Hombo and Donoghue, 1999, 2000) have found that the asymptotic distribution functioned extremely well, even with samples as small as 1000 examinees. Both $Q - Q$ plots and Type I error rates indicated very good agreement between the asymptotic distribution and the observed values. Moreover, the measure has good power to detect misfit when it was present in items (Hombo and Donoghue, 2001).

The difference between the second approach and the first one is that Q^*_{DM} is based on the distribution of ability, at each quadrature point. The term "pseudocount" by

Donoghue, McClellan and Orange(e.g., 2004) refers to the fact that real counts of the number of examinee proficiency estimates falling with an interval on the scale are not used. Rather, counts are estimated from the sum of posterior distributions. Peudocounts are the basic building blocks for the item fit measure Q^*_{DM}. *Pseudocounts* of examinees at a given quadrature point are computed by summing over the posterior expectation (pseudocounts) of an M-category item for score level k and proficiency θ level q. Then Q^*_{DM} is defined as

$$Q^*_{DM} = \sum_{q=1}^{Q} \sum_{k=0}^{M} \frac{(O_{kq} - E_{kq})^2}{E_{kq}}. \tag{1.1}$$

Here O represents the observed response counts and E represents the expected response counts. Assuming that item parameters are known, Q^*_{DM} has been shown to be asymptotically distributed as a quadrature form of normal variables (Donoghue and Hombo, 1999). This distribution is represented as the sum of independent $\chi^2_{(1)}$ variates (e.g., Johnson and Kotz, 1970). $Q^*_{DM} \sim \sum_{i=1}^{m} \lambda_i \chi^2_{(1)}$, where $\lambda_i, \forall i = 1, 2, \cdots, m$, are the non-zero eigenvalues of matrix $L'\Sigma L$, L is a special form of matrix with dimension $2Q \times Q$ (Q is the number of quadrature point used in the computation) for dichotomous items, and Σ is the covariance matrix of the pseudocounts (Donoghue, McClellan, and Oranje, 2004). A routine by Davies (1980) can be used to evaluate this probability.

However, further work is needed to establish the utility of the result in practical testing situations. Hombo and Donoghue (1999, 2000) examined some possible limiting factors, including potentially prohibitive sample size requirements to achieving sampling distribution properties approaching those of the asymptotic distribution. A

major limitation to practical application of the findings is the computational burden required to compute the asymptotic distribution Q_{DM}. The computation requires the evaluation of all possible item response patterns—2^J for a test of J dichotomous items, for example. For short-moderate length tests ($10 - 15$ items) the number for patterns (1024-32768) is manageable. For tests of 20 items, the evaluation of slightly over one million response patterns per item begins to become burdensome.

1.5 Approximation by Observed Covariance Among Pseudocounts

The work for the asymptotic distribution for the item fit measure Q^*_{DM} represents a major advance along this line of research. To avoid evaluating all possible response patterns for calculating the covariance matrix of pseudocounts and thus making applications possible to operational research, Donoghue, McClellan, and Oranje (2004) propose a consistent estimator $\mathbf{S}$ for the covariance matrix $\mathbf{\Sigma}$ and the true asymptotical distribution is approximated by the observed matrix of interrelations among pseudocounts. To understand and construct the matrix S, consider the joint probability consisting of positive values $p(U = u_i, \theta_q)$ and 0 for $p(U \neq u_i, \theta_q)$ for dichotomous item i and given response u_i and any quadrature point θ_q, $\forall q = 1, 2, \cdots, Q$, and $i = 1, 2, \cdots, J + 1$. Then S can be seen as a simple covariance matrix with every examinee contributing to all of the $2Q$ quadrature points. The matrix $\mathbf{S}$ is a consistent estimator of $\mathbf{\Sigma}$. Therefore, a natural idea is to use the data-based estimator $L'SL$ in place of$L'\Sigma L$. Because Q_{DM} is an asymptotic result, for very large N (approaching

infinity) is arbitrarily close to Σ and intuitively should yield the correct estimate of Q_{DM}.

Indeed, the use of the observed matrix of interrelation among pseudocounts yields the hoped-for accuracy and simplicity on computation, and the approximation of Q_{DM} based on the observed matrix of interrelations among the pseudocounts opens up the possibility of operationally feasible and theoretically defensible statistical test of item misfit. Results from Li, Donoghue, and McClellan (2005) demonstrate how accurate the approximation is in relative to the asymptotic distributions across three different sample sizes. The results from simulation studies show that the approximation works extremely well for many situations. The cumulative probability, mean, and variance are very close between the true and approximation values. These results can also be generalized to the case of polytomous items, as in Donoghue and Hombo (2001a) when item parameters are known constrants.

However, the asymptotic distribution of Q^*_{DM} was derived under the assumption that the item parameters are fixed and known. When the item parameters are data-based estimates, the theoretic results of Donoghue and Hombo (1999) do not hold. Several studies (Donoghue and Isham, 1998; Hombo and Donoghue, 1999; Donoghue and Hombo, 2001ab; Stone and Zhang, 2002) have repeatedly found that, when item parameters are data-based estimates, Type I error rates from Q_{DM} are much too conservative, and that distribution of the Q^*_{DM} statistic is stochastically smaller than Q_{DM}. This study is an attempt to overcome the disadvantage of working with the item parameters by reformulating the measure of item fit based on pseudocounts.

1.6 Reformulating the Item Fit Measure Q^*_{DM}

The form of Q^*_{DM} defined as in 1.1 is a Person-type measure for goodness-of-fit. Donoghue and Hombo (e.g., 1999) suggest that the expectation of the pseudocounts can be found through binomial approximation. That is, the expectation of pseudocounts is a product of total pseudocounts and the hypothetical item response function at certain levels of quadrature points (please refer to the first section of Chapter 2). The asymptotic distribution of Q^*_{DM} can be shown through a Taylor expansion of the fit statistic. As sample size increases, the asymptotic distribution for the second order Taylor expansion of Q^*_{DM} converges to the true asymptotic distribution of Q^*_{DM}.

The idea of reformulating Q^*_{DM} is to simply replace the expectation of pseudocounts by its theoretic expectation under null hypothesis. The reformulated version of the statistic Q^*_{DM} allows researchers to derive the true asymptotic reference distribution for Q^*_{DM} and to extend the results for data-based item parameter estimates.

Chapter 2

Item Fit Analysis Based on Pseudocounts

The item fit measure Q^*_{DM} by Donoghue and McClellan (e.g., 2004, 2003b, 2003a, 2001b, 2001a, 1999) is similar in form to a Pearson χ^2. However, as noted before, the distribution of Q^*_{DM} is not χ^2, but a quadratic function of normal variates. This chapter first introduces the basic concept of pseudocounts, on which the measures of item fit (i.e., Q^*_{DM}) are based. Next the reformulation of D^*_{DM} will be discussed with the help of the fundamental concept of pseudocounts. Then the asymptotic distribution of the reformulated measure of item fit will be derived in a different way. Finally, the observed interrelations among pseudocounts are examined to obtain a consistent estimator of the true covariance matrix among pseudocounts.

2.1 Definitions and Notations

Let θ_q be the discrete proficiency at quadrature point q, $w(\theta_q) = w_q$ be the density of θ, i.e., $P(\theta = \theta_q) = w_q$. The prior w will often be chosen to approximate a continuous distribution, such as $N(\mu, \sigma^2)$. Denote U as a random variable representing the

response for dichotomously scored studied item. In study of item level model fit, test items are classified into two groups—the studied item (only one item) and the remaining items (containing J items). Thus the total number of items in the test is $J+1$.

Let $f_{q1} = f(\theta_{q1})$ be the item response function for the studied item, i.e., $P(U = 1|\theta = \theta_q)$. Let N be the sample size or number of examinees, and t index patterns of responses to the remaining J items $\mathbf{Y}$ on a test. For the dichotomous items, $t = 1, \cdots, T = 2^J$. Let n_{tk} be the number of examinees who got score pattern $(U = k, Y = y_t)$. Suppose $\hat{\pi}$ is the vector of observed proportions for the sample response pattern $(U = k, Y = y_t)$. Then $\hat{\pi}_{tk} = n_{tk}/N$, and $l_{tq} = P(Y = y_t|\theta = \theta_q)$, where k represents the category for the studied items (e.g., for dichotomous case, $k = 0, 1$), and l_{tq} is the likelihood function of the remaining item response pattern $(Y = y_t)$ at quadrature point q. Denote π_{tk} the model-based prediction of the probability of response pattern $(U = k, Y = Y_t)$, or the marginal probability of $(U = k, Y = Y_t)$. For dichotomous case (i.e., $k = 0, 1$), it is easy to see that

$$\begin{aligned} \pi_{t1} &= P(U = 1, Y = y_t) \\ &= \sum_{q=1}^{Q} w_q f_{q1} l_{tq}. \end{aligned}$$

Similarly, $\pi_{t0} = \sum_{q=1}^{Q} w_q(1 - f_{q1}) l_{tq}$. Let p_{tq}^k be the posterior of θ at quadrature point $\theta = \theta_q$ given response pattern $(U = k, Y = y_t)$. Then,

$$\begin{aligned} p_{tq}^k &= P(\theta = \theta_q|U = k, Y = y_t) \\ &= \frac{w_q f_{qk} l_{tq}}{\pi_{tk}}. \end{aligned}$$

In dichotomous case, the posterior distribution for $\theta = \theta_q$ given the response pattern $(U = 1, Y = y_t)$ is $p_{tq}^1 = \frac{w_q f_{q1} l_{tq}}{\pi_{t1}}$, and $p_{tq}^0 = \frac{w_q (1-f_{q1}) l_{tq}}{\pi_{t0}}$ given response pattern $(U = 0, Y = y_t)$. The posterior distributions provide the best information about the distribution of examinees' proficiency levels. Thus, it is the posterior distribution of proficiency rather than the proficiency point estimates that are used for assessing model-data fit on the item level in this regard.

Define pseudocount, s_{q1}, to response U for the studied item at quadrature point θ_q as the sum of the posteriors over all response patterns $P(\theta = \theta_q | U, Y = y_t), \forall t = 1, 2, \cdots, T$. For example, the pseudocount to the correct response for the studied item at quadrature point θ_q.

$$\begin{aligned} s_{q1} &= \sum_{t=1}^{T} n_{t1} p_{tq}^1 \\ &= w_q f_{q1} \sum_{t=1}^{T} \frac{n_{t1} l_{tq}}{\pi_{t1}}. \end{aligned}$$

Here T is the number of all possible response patterns for the remaining items in the test. In a similar fashion, define s_{q0} as $s_{q0} = w_q(1 - f_{1q}) \sum_{t=1}^{T} \frac{n_{t0} l_{tq}}{\pi_{t0}}$. Denote $\bar{s}_q = s_{q1} + s_{q0}$. s_q is the total pseudocount at quadrature point θ_q, $\forall q = 1, 2, \cdots, Q$. Q is the total number of quadrature points (designated in the study, in this case 41, ranging from -4 to 4).

Now consider the following vectors in the dichotomous case:

$\mathbf{n}^T = (n_{11}, n_{21}, \cdots, n_{T1}, n_{10}, n_{20}, \cdots, n_{T0})$,

$\hat{\pi}^T = (\hat{\pi}_{11}, \hat{\pi}_{21}, \cdots, \hat{\pi}_{T1}, \hat{\pi}_{10}, \hat{\pi}_{20}, \cdots, \hat{\pi}_{T0}) = \mathbf{n}/N$,

$\pi^T = (\pi_{11}, \pi_{21}, \cdots, \pi_{T1}, \pi_{10}, \pi_{20}, ..., \pi_{T0})$, the model-based probabilities,

$\mathbf{s}^T = (s_{11}, s_{21}, ..., s_{Q1}, s_{10}, s_{20}, s_{Q0})$, observed pseudocounts,

$\bar{\mathbf{s}}^T = (\bar{s}_1, \bar{s}_2, \cdots, \bar{s}_Q)$.

The vector $\mathbf{n}$ describes the frequencies of all possible patterns of the response data for $J+1$ items in a test. That is, $\mathbf{n}$ contains the frequencies of the mutually exclusive response patterns from the sample data. If N examinees are available, then $\sum_{t=0}^{T}(n_{t1}+n_{t0}) = N$. The model-based probability of the tth pattern of the remaining items and correct response on the studied item (i.e., $(U=1, Y=y_t)$) is $\pi_{t1} = P(U=1, Y=Y_t), \forall t = 1, 2, ..., T$. Similarly, the probability of observing response $(U=0, Y=y_i)$ is $\pi_{t0} = P(U=0, Y=y_t), \forall t = 1, 2, ..., T$.

For the convenience of studying the statistical properties of pseudocounts, two posterior matrices $\mathbf{P}$ and $\tilde{\mathbf{P}}$ are constructed. $\mathbf{P}$ is a matrix consisting of all posterior and having dimension of $2T$ by $2Q$. That is,

$$\mathbf{P_{2T\times 2Q}} = \begin{pmatrix} p^1_{11} & p^1_{12} & ... & p^1_{1Q} & 0 & 0 & ... & 0 \\ p^1_{21} & p^1_{22} & \cdots & p^1_{2Q} & 0 & 0 & \cdots & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ p^1_{T1} & p^1_{T2} & \cdots & p^1_{TQ} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & p^0_{11} & p^0_{12} & \cdots & p^0_{1Q} \\ 0 & 0 & \cdots & 0 & p^0_{21} & p^0_{22} & \cdots & p^0_{2Q} \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & \cdots & 0 & p^0_{T1} & p^0_{T2} & \cdots & p^0_{TQ} \end{pmatrix}.$$

The matrix $\tilde{\mathbf{P}}$ is a $2T \times Q$ matrix defined as

$$\tilde{\mathbf{P}} = \begin{pmatrix} p^1_{11} & p^1_{12} & \cdots & p^1_{1Q} \\ p^1_{21} & p^1_{22} & \cdots & p^1_{2Q} \\ \cdots & \cdots & \cdots & \cdots \\ p^1_{T1} & p^1_{T2} & \cdots & p^1_{TQ} \\ p^0_{11} & p^0_{12} & \cdots & p^0_{1Q} \\ p^0_{21} & p^0_{22} & \cdots & p^0_{2Q} \\ \cdots & \cdots & \cdots & \cdots \\ p^0_{T1} & p^0_{T2} & \cdots & p^0_{TQ} \end{pmatrix}.$$

With the matrix notation, the pesudocount vector $\mathbf{s}$ can be expressed as $\mathbf{s} = \mathbf{P}'\mathbf{n}$ and $\bar{\mathbf{s}} = \tilde{\mathbf{P}}'\mathbf{n}$. The matrix $\mathbf{P}$ can be written as column form $\mathbf{P} = (P_1^1, P_2^1, \cdots, P_Q^1, P_1^0, P_2^0, \cdots, P_Q^0)$, where $P_q^j, \forall q = 1, 2, \cdots, Q$ and $j = 1, 0$, denotes the column in the matrix $\mathbf{P}$ corresponding to the posteriors at quadrature point θ_q with response $U = j$ for the studied item. Then $s_{qj} = P_q^{jT}\mathbf{n}$. Similarly, write the matrix $\tilde{\mathbf{P}}$ as $\tilde{\mathbf{P}} = (\tilde{P}_1, \tilde{P}_2, \cdots, \tilde{P}_Q)$, where $\tilde{P}_q$ represents the qth column in the matrix $\tilde{\mathbf{P}}, \forall q = 1, 2, \cdots, Q$. Then $\bar{s}_q = \tilde{P}_q^T\mathbf{n}$.

The pseudocount vector $\mathbf{s}$ or $\bar{\mathbf{s}}$ can be considered as a random vector since it is a linear function of the frequency vector $\mathbf{n}$, which follows multinomial distribution with probability vector π, denoted as $\mathbf{n} \sim M_{2T}(N, \pi)$ with $\sum_{t=1}^{T}(\pi_{t1} + \pi_{t0}) = 1$.

To establish the results regarding the asymptotic distribution of the pseudo-counts vector $\mathbf{s}$, the following two vectors are useful:

$$\mathbf{v}^T = \left(\frac{n_{11}-N\pi_{11}}{\sqrt{N\pi_{11}}}, \frac{n_{21}-N\pi_{21}}{\sqrt{N\pi_{21}}}, \ldots, \frac{n_{T1}-N\pi_{T1}}{\sqrt{N\pi_{T1}}}, \frac{n_{10}-N\pi_{10}}{\sqrt{N\pi_{10}}}, \frac{n_{20}-N\pi_{20}}{\sqrt{N\pi_{20}}}, \ldots, \frac{n_{T0}-N\pi_{T0}}{\sqrt{N\pi_{T0}}}\right),$$

$$\varphi^T = \left(\sqrt{\pi_{11}}, \sqrt{\pi_{21}}, \ldots, \sqrt{\pi_{T1}}, \sqrt{\pi_{10}}, \sqrt{\pi_{20}}, \ldots, \sqrt{\pi_{T0}}\right).$$

The object is to study the properties regarding the pseudocounts, which are a linear combination of the observed frequency vector $\mathbf{n}$ of response patterns.

2.2 Asymptotic Distributions of Pseudocounts

Before showing the theorems regarding the pseudocounts, define a matrix $\mathbf{B}$ with fixed elements, $\mathbf{B} = \mathbf{D}_\pi^{\frac{1}{2}}\mathbf{P}$, where $\mathbf{P}$ is the matrix of posteriors defined as before, $\mathbf{D}_\pi^{\frac{1}{2}}$ is a diagonal matrix with square root of the model-based prediction vector π as its

diagonal entries.

$$\mathbf{D}_{\pi}^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\pi_{11}} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \sqrt{\pi_{21}} & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \sqrt{\pi_{T1}} & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sqrt{\pi_{10}} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \sqrt{\pi_{20}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \sqrt{\pi_{T0}} \end{pmatrix}.$$

If item parameters are all known constants, so are each component in the posterior matrix $\mathbf{P}$ and each element in the diagonal matrix $\mathbf{D}_{\pi}^{\frac{1}{2}}$. Simply put, the product matrix $\mathbf{B}$ has entries of fixed values. Denote each column of $\mathbf{B}$ as $\mathbf{b_q^1}$ or $\mathbf{b_q^0}, \forall q = 1, 2, \cdots, Q$. Then $\mathbf{B}$ can be expressed as $\mathbf{B} = (\mathbf{b_1^1}, \mathbf{b_2^1}, ..., \mathbf{b_Q^1}, \mathbf{b_1^0}, \mathbf{b_2^0}, \cdots, \mathbf{b_Q^0})$. $\mathbf{b_q^1}$ or $\mathbf{b_q^0}$ is a fixed vector with dimension of $2T$, and $\mathbf{b}_q^1 = \mathbf{D}_{\pi}^{\frac{1}{2}} P_q^1$, or $\mathbf{b}_q^0 = \mathbf{D}_{\pi}^{\frac{1}{2}} P_q^0$.

theorem 2.2.1 (Marginal Distribution of Pseudocounts) *The asymptotic distribution of* $\sqrt{N}(\frac{s_{qj}}{N} - P_q^{jT}\pi)$ *for each element* s_{qj} *in the pseudocount vector* $\mathbf{s}$ *defined as above is normal with mean 0 and variance* $P_q^{j'}(\mathbf{D}_{\pi} - \pi\pi')P_q^j, \forall q = 1, 2, \cdots, Q$ *and* $j = 0, 1$.

Proof: Let the vectors $\mathbf{v}$, $\mathbf{b_q^1}$ or $\mathbf{b_q^0}$, $\forall q = 1, 2, \cdots, Q$ be defined as above. Then the asymptotic distribution of the linear function of $\mathbf{b_q^1 v}$ or $\mathbf{b_q^0 v}$ is normal with mean 0 and variance $\mathbf{b_q^1 b_q^{1'}} - (\mathbf{b_q^1}\varphi)^2 = \mathbf{b_q^{1'}}(\mathbf{I} - \varphi\varphi')\mathbf{b_q^1}$, or $\mathbf{b_q^0 b_q^{0'}} - (\mathbf{b_q^0}\varphi)^2 = \mathbf{b_q^{0'}}(\mathbf{I} - \varphi\varphi')\mathbf{b_q^0}$,

respectively (p383, Rao, 1973). Therefore,

$$\begin{aligned}
\mathbf{b}_q^{j'}\mathbf{v} &= \frac{1}{\sqrt{N}}\mathbf{b}_q^{j'}\mathbf{D}_\pi^{-\frac{1}{2}}(\mathbf{n}-N\pi) \\
&= \frac{1}{\sqrt{N}}P_q^{j^T}(\mathbf{n}-N\pi) \\
&= \frac{1}{\sqrt{N}}\sum_{t=1}^{T}(n_{tj}p_{tq}^j - N\pi_{tj}p_{tq}^j) \\
&= \sqrt{N}(\frac{s_{q1}}{N} - \sum_{t=1}^{T}\pi_{tj}p_{tq}^j) \\
&= \sqrt{N}(\frac{s_{q1}}{N} - P_q^{j^T}\pi).
\end{aligned}$$

$Var(\mathbf{b}_q^{j'}\mathbf{v}) = P_q^{j^T}\mathbf{D}_\pi^{\frac{1}{2}}(\mathbf{I}-\varphi\varphi')P_q^j = P_q^{j'}(\mathbf{D}_\pi - \pi\pi')P_q^j$. Hence the theorem.

The expectation and variance of pseudocounts, Es_{qj} and $Var(s_{qj})$ respectively, can be found easily by $Es_{qj} = EP_q^{j^T}\mathbf{n} = NP_q^{j^T}\pi = Nw_qf_{q1}$ and $Var(s_{qj}) = P_q^{j^T}, Var(\mathbf{n})P_q^j = NP_q^{j^T}(\mathbf{D}_\pi - \pi\pi')P_q^j), \forall q = 1,2,\cdots,Q$ and $j = 1,0$. It can be seen from the theorem that each pseudocount is a random variable and asymptotically distributed as normal. Or the sequence of the pseudocounts s_{11}, $\sqrt{N}(\frac{s_{11}}{N} - P_1^{1^T}\pi) \sim N(0, P_1^{1'}(\mathbf{D}_\pi - \pi\pi')P_1^1)$ asymptotically as $N \to \infty$, where vector P_1^1 can be written as $P_1^{1'} = (p_{11}^1, p_{21}^1, \cdots, p_{T1}^1, 0, 0, \cdots, 0)$. And $E(s_{11}) = N\sum_{t=1}^{T}\pi_{t1}p_{t1}^1$, and $Var(s_{11}) = N(\sum_{t=1}^{T}{p_{t1}^1}^2\pi_{t1} - (\sum_{t=1}^{T}p_{t1}^1\pi_{t1})^2)$.

In the same way, it can be shown that the marginal distribution of the total pseudocount $\bar{s}_q$ at the quadrature point θ_Q is also asymptotically normally distributed with mean 0 and variance $\tilde{P}_q'(\mathbf{D}_\pi - \pi\pi')\tilde{P}_q$, i.e., $\sqrt{N}(\frac{\bar{s}_q}{N} - \tilde{P}_q^T\pi) \sim N(0, \tilde{P}_q'(\mathbf{D}_\pi - \pi\pi')\tilde{P}_q), \forall q = 1,2,\cdots,Q$, where $\tilde{P}_q$ is the qth column in the matrix $\tilde{\mathbf{P}}$. Note that the expectation of $\bar{s}_q$ is $E\bar{s}_q = N\tilde{P}_q'\pi = Nw_q$. Interestingly, notice that $E\bar{s}_q$ does not depend on the hypothetical models. $E\bar{s}_q$ only depends on the quadrature approximation

w_q.

theorem 2.2.2 Joint Distribution of Pseudocounts *The asymptotic distribution of* $\sqrt{N}(\frac{\mathbf{s}}{N} - \mathbf{P}'\pi)$ *for the pseudocounts vector* $\mathbf{s}$ *is multivariate normal with mean vector 0 and dispersion matrix* $\mathbf{P}'(\mathbf{D}_\pi - \pi\pi')\mathbf{P}$, *where* $\mathbf{P}, \mathbf{D}_\pi$, *and* π, *are defined as above.*

Proof: The asymptotic distribution of the $2Q$ linear functions $\mathbf{B}'\mathbf{v}$, where $\mathbf{B}$ is a $2T$ by $2Q$ matrix of rank $2Q-2$ defined as above, is multivariate normal with mean vector zero and dispersion matrix $\mathbf{B}'(\mathbf{I} - \varphi\varphi')\mathbf{B}$ (p383, Rao, 1973). It is easy to see that

$$\begin{aligned}\mathbf{B}'\mathbf{v} &= (\mathbf{b_1^1}, \mathbf{b_2^1}, \cdots, \mathbf{b_Q^1}, \mathbf{b_1^0}, \mathbf{b_2^0}, \cdots, \mathbf{b_Q^0})^{\mathbf{T}}\mathbf{v} \\ &= \frac{S}{\sqrt{N}} - \sqrt{N}\mathbf{B}'\varphi.\end{aligned}$$

After a little algebra, it can be shown that the pseudocounts vector is asymptotically multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{B}'(\mathbf{I}-\varphi\varphi')\mathbf{B}$, i.e., $\sqrt{N}(\frac{\mathbf{s}}{N}-\mathbf{P}'\pi) \sim \mathbf{N_{2Q}}(\mathbf{0}, \mathbf{P}'(\mathbf{D}_\pi-\pi\pi')\mathbf{P})$ asymptotically as $N \to \infty$. Similarly, the asymptotical distribution of $\sqrt{N}(\frac{\bar{\mathbf{s}}_q}{N} - \tilde{\mathbf{P}}'\pi)$for total pseudocounts vector $\bar{\mathbf{s}}$ is $N_Q(\mathbf{0}, \tilde{\mathbf{P}}'(\mathbf{D}_\pi - \pi\pi')\tilde{\mathbf{P}})$ as $N \to \infty$.

Now one can see why pseudocounts contain essential information for assessing the degree of item fit. They are the sum of posterior distributions across all possible response patterns and over all examinees. The posterior probability of proficiency, instead of the count of grouped proficiency estimates themselves, provide the best information for evaluating the degree of model-data fit. The proportions of pseudocounts $\mathbf{s}$ over the total number of examinees N can give empirical values that can be

compared to IRT model predicted values. A measure of the correspondences between the empirical and predicted values represents the degree of adequacy of model-data fit at the item level. However, it is often difficult or impossible to judge from the plots whether the differences between the empirical values based on pseudocounts and the model based predicted values. A statistical significance test is very desirable. The following section is to reformulate Q^*_{DM} and find out its reference distribution based on pseudocounts.

2.3 The Asymptotic Distribution of the Item Fit Measure Q^*_{DM}

The statistic Q^*_{DM} suggested by Donoghue and McClellan (e.g., 2003) is defined through binomial approximating the expectation of pseudocounts as

$$\begin{aligned} Q^*_{DM} &= \sum_{q=1}^{Q} \left(\frac{(s_{q1} - Es_{q1})^2}{Es_{q1}} + \frac{(s_{q0} - Es_{q0})^2}{Es_{q0}} \right) \\ &= \sum_{q=1}^{Q} \left(\frac{(s_{q1} - f_{q1}s_q)^2}{f_{q1}s_q} + \frac{(s_{q0} - f_{q0}s_q)^2}{f_{q0}s_q} \right) \\ &= \sum_{q=1}^{Q} \frac{(s_{q1} - f_{q1}s_q)^2}{f_{q1}(1 - f_{q1})s_q}. \end{aligned}$$

Donoghue and Hombo (2003b) expand the above expression of Q^*_{DM} about $\hat{\pi} = \pi$ as a Taylor series to derive the that the asymptotic distribution of the measure Q^*_{DM} is asymptotically a quadratic form of normal variables:

$$Q^*_{DM}(\hat{\pi}) = \sqrt{N}(\hat{\pi} - \pi)' \mathbf{C} (\hat{\pi} - \pi)\sqrt{N} + o(N^{-\frac{1}{2}})$$

The matrix **C** is the same as that in Donoghue, McClellan, and Oranje (2004, p 10). That is,

$$\mathbf{C} = \sum_{q=1}^{Q} w_q \frac{(\mathbf{v_{q1}}^* - f_{q1}\mathbf{v_{q2}^*})(\mathbf{v_{q1}^*} - f_{q1}\mathbf{v_{q2}^*})}{f_{q1}(1-f_{q1})},$$

where $\mathbf{v_{q1}^*}' = \left(\frac{f_{q1}l_{1q}}{\pi_{11}}, \frac{f_{q1}l_{2q}}{\pi_{21}}, \cdots, \frac{f_{q1}l_{Tq}}{\pi_{T1}}, 0, \cdots, 0\right)$, and $\mathbf{v_{q2}^*}' = \left(\frac{f_{q1}l_{1q}}{\pi_{11}}, \frac{f_{q1}l_{2q}}{\pi_{21}}, \cdots, \frac{f_{q1}l_{Tq}}{\pi_{T1}}, \frac{(1-f_{q1})l_{1q}}{\pi_{10}}, \frac{(1-f_{q1})l_{2q}}{\pi_{20}}, \cdots, \frac{(1-f_{q1})l_{Tq}}{\pi_{T0}}, \right)$.

2.3.1 Reformulated Q_{DM}^* and Its Asymptotic distribution

In this study, also define Q_{DM}^* as Pearson χ^2-like statistic. That is,

$$Q_{DM}^* = \sum_{q=1}^{Q} \left(\frac{(s_{q1} - Es_{q1})^2}{Es_{q1}} + \frac{(s_{q0} - Es_{q0})^2}{Es_{q0}} \right).$$

As previously defined, s_{q1} or s_{q0} is the pseudocount at quadrature point $\theta_q, \forall q = 1, 2, \cdots, Q$. Es_{q1} or Es_{q0} denote the corresponding expectations. First simplify the expression of the expectation of s_{q1} and $s_{q0}, \forall q = 1, 2, \cdots, Q$. Notice that the expectation of the pseudocount $Es_{qj} = N\mathbf{P_q'}\pi$ for $j = 0, 1$ can be expressed as

$$\begin{aligned} Es_{qj} &= E\left(N w_q f_{qj} \sum_{t=1}^{T} \frac{\hat{\pi}_{tj} l_{tq}}{\pi_{tj}} \right) \\ &= N w_q f_{qj} \sum_{t=1}^{T} l_{tq} \\ &= N w_q f_{qj}. \end{aligned}$$

That is, $Es_{q1} = N w_q f_{q1}$, and $Es_{q0} = N w_q f_{q0} = N w_q (1 - f_{q1})$. Therefore, the expectation of the pseudocounts vector **s** is $E\mathbf{s} = N(w_1 f_{11}, w_2 f_{21}, \cdots, w_Q f_{Q1}, w_1 f_{10}, w_2 f_{20}, \cdots, w_Q f_{Q0})^T$. The expression of $E\mathbf{s}$ is the same as that derived from the theorem on joint distribution of the pseudocounts vector.

Now turn to look at the asymptotic distribution of the reformulated Q^*_{DM}. Let $\mathbf{D_{ES}}$ be a diagonal matrix with the expectation of the pseudocounts as its diagonal elements. Obviously, $\mathbf{D_{ES}^{-1}}$ is a $2Q$ by $2Q$ matrix, and $\mathbf{D_{ES}^{-1}} = \mathbf{D_{ES}^{-\frac{1}{2}}}\mathbf{D_{ES}^{-\frac{1}{2}}}$, where $\mathbf{D_{ES}^{-\frac{1}{2}}}$ can be expressed as

$$\mathbf{D_{ES}}^{-\frac{1}{2}} = \frac{1}{\sqrt{N}} \begin{pmatrix} \frac{1}{\sqrt{w_1 f_{11}}} & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \frac{1}{\sqrt{w_2 f_{21}}} & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \frac{1}{\sqrt{w_Q f_{Q1}}} & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{w_1 f_{10}}} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \frac{1}{\sqrt{w_2 f_{20}}} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & \frac{1}{\sqrt{w_Q f_{Q0}}} \end{pmatrix}.$$

With the matrix $\mathbf{D_{ES}}^{-\frac{1}{2}}$, the Q^*_{DM} can be further simplified by

$$\begin{aligned} Q^*_{DM} &= \sum_{q=1}^{Q} \left(\frac{(s_{q1} - Es_{q1})^2}{Es_{q1}} + \frac{(s_{q0} - Es_{q0})^2}{Es_{q0}} \right) \\ &= \|\mathbf{D_{Es}}^{-\frac{1}{2}}(\mathbf{s} - \mathbf{Es})\|^2 \\ &= (\mathbf{s} - \mathbf{Es})' \mathbf{D_{Es}}^{-1} (\mathbf{s} - \mathbf{Es}) \\ &= (\mathbf{P}'\mathbf{n} - \mathbf{N}\mathbf{P}'\pi)' \mathbf{D_{ES}}^{-1} (\mathbf{P}'\mathbf{n} - \mathbf{N}\mathbf{P}'\pi) \\ &= (\mathbf{n} - \mathbf{N}\pi)' \mathbf{P} \mathbf{D_{Es}^{-1}} \mathbf{P}' (\mathbf{n} - \mathbf{N}\pi) \\ &= \sqrt{N}(\hat{\pi} - \pi)' \mathbf{N} \mathbf{P} \mathbf{D_{ES}^{-1}} \mathbf{P}' \sqrt{\mathbf{N}} (\hat{\pi} - \pi). \end{aligned}$$

As it is known, $\sqrt{N}(\hat{\mathbf{p}} - \pi)$ are asymptotically distributed as multivariate normal variates with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{G} = \mathbf{D}_\pi - \pi\pi'$ (e.g., p470, Bishop, Fienberg, and Holland, 1975). Thus, Q^*_{DM} is asymptotically a quadratic function of normal variables. Obviously, the matrix $N\mathbf{P}\mathbf{D_{ES}^{-1}}\mathbf{P}'$ is nonnegative definite since all of the diagonal components in the matrix $\mathbf{D_{ES}}$ are nonnegative. Following Sta-

pleton's (1995, p65) expression for quadratic form by denoting $\mathbf{y} = \sqrt{\mathbf{N}}(\hat{\mathbf{p}} - \pi) \sim \mathbf{N_{2T}}(\mathbf{0}, \mathbf{G})$, and the nonnegative definite matrix $\mathbf{A} = \mathbf{NPD_{ES}^{-1}P'}$, let $\mathbf{G}^{\frac{1}{2}}$ be the unique symmetric square root of $\mathbf{G}$, and let $\mathbf{G}^{-\frac{1}{2}}$ be its inverse. Thus $Q_{DM}^* = (\mathbf{y'G}^{-\frac{1}{2}})(\mathbf{G}^{\frac{1}{2}}\mathbf{AG}^{\frac{1}{2}})(\mathbf{G}^{-\frac{1}{2}}\mathbf{y}) = \mathbf{z'Cz}$, where $\mathbf{z} = \mathbf{G}^{-\frac{1}{2}}\mathbf{y}$ and $\mathbf{C} = \mathbf{G}^{\frac{1}{2}}\mathbf{AG}^{\frac{1}{2}}$. Then $Var(\mathbf{z}) = \mathbf{G}^{-\frac{1}{2}}\mathbf{GG}^{-\frac{1}{2}}$ is an identity matrix of $2T \times 2T$, so that $\mathbf{z} \sim \mathbf{N_{2T}}(\mathbf{0}, \mathbf{I})$. Let $\mathbf{C} = \mathbf{T\Lambda T'}$ be the spectral decomposition of $\mathbf{C}$. Then $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{C}$, and $\mathbf{T}$ is the $2T \times 2T$ matrix whose columns are the corresponding eigenvectors of $\mathbf{C}$, and $\mathbf{T}$ is an orthogonal matrix. Hence, $Q_{DM}^* = \mathbf{z'T\Lambda T'z} = (\mathbf{T'z})'\mathbf{\Lambda}(\mathbf{T'z}) = \omega'\mathbf{\Lambda}\omega$ for $\omega = \mathbf{T'z}$. $Var(\omega) = \mathbf{T'IT} = \mathbf{I_{2T \times 2T}}$, and $\omega \sim \mathbf{N_{2T}}(\mathbf{0}, \mathbf{I})$. Denoting the eigenvalues of $\mathbf{C}$ by $\lambda_1, \lambda_2, \cdots, \lambda_{2T}$, $Q_{DM}^* = \sum_{i=1}^{2T} \lambda_i \omega_i^2$, where $\omega' = (\omega_1, \omega_2, \cdots, \omega_{2T})$. Therefore, Q_{DM}^* is a linear combination with coefficients $\lambda_1, \lambda_2, \cdots, \lambda_{2T}$ of independent χ_1^2 random variables. The coefficients $\lambda_1, \lambda_2, \cdots, \lambda_{2T}$ are the eigenvalues of $N\mathbf{G'}^{\frac{1}{2}}\mathbf{PD_{ES}^{-1}P'G'}^{\frac{1}{2}}$, and also the eigenvalues of $\mathbf{NPD_{ES}^{-1}P'G}$ or of $\mathbf{NGPD_{ES}^{-1}P'}$. By theorem 2.2.1 (Stapletone, 1995, p51), the expectation of Q_{DM}^* is $E(Q_{DM}^*) = trace(\mathbf{AG}) = trace(N\mathbf{PD_{ES}^{-1}P'G})$.

The asymptotic distribution of Q_{DM}^* can further be simplified as the reduced sum of independent $\chi_{(1)}^2$ variates (e.g., Johnson and Kotz, 1970). That is, $Q_{DM}^* \sim \sum_{i=1}^{m} \lambda_i \chi_i^2$, where λ_i are the m non-zero eigenvalues of the $2T \times 2T$ matrix $N\mathbf{PD_{ES}^{-1}P'G}$. The non-zero eigenvalues from matrix $N\mathbf{PD_{ES}^{-1}P'G}$ is equivalent to the non-zero eigenvalues from matrix $\mathbf{L'GL} = \mathbf{ND_{Es}^{-\frac{1}{2}}P'(D_\pi - \pi\pi')PD_{Es}^{-\frac{1}{2}}}$ for $\mathbf{L} = \mathbf{P'D_{ES}^{-\frac{1}{2}}}$. It can be easily seen by letting ν be a non-zero vector (with dimension of $2Q$) and scalar λ. Then by defining equation $N\mathbf{PD_{ES}^{-1}P'G}\nu = \lambda\nu$, $NLL'\mathbf{G}\nu = \lambda\nu$. This implies

that $NL'\mathbf{G}\nu = \lambda L^{\dagger}\nu$, where $L^{\dagger}$ represents the generalized inverse of matrix L. And $NL'\mathbf{G}L(L^{\dagger}\nu) = \lambda(L^{\dagger}\nu)$. Hence the result. A routine by Davies (1980) can be used to evaluate this probability distribution. Now state this result about the asymptotic distribution of Q^{*}_{DM} in the following theorem.

theorem 2.3.1 Asymptotic Distribution of Q^{*}_{DM} *The Pearson χ^2-like measure of item fit Q^{*}_{DM} defined as above is a quadratic function of random variables with mean vector* $\mathbf{0}$ *and covariance matrix* $\mathbf{D}_{\pi} - \pi\pi'$.

Take a close look at the covariance matrix $N\mathbf{D}_{\mathbf{Es}}^{-\frac{1}{2}}\mathbf{P}'(\mathbf{D}_{\pi} - \pi\pi')\mathbf{P}\mathbf{D}_{\mathbf{Es}}^{-\frac{1}{2}}$. Denote this matrix product as $\mathbf{A}$ (i.e., $\mathbf{A} = N\mathbf{D}_{\mathbf{Es}}^{-\frac{1}{2}}\mathbf{P}'(\mathbf{D}_{\pi} - \pi\pi')\mathbf{P}\mathbf{D}_{\mathbf{Es}}^{-\frac{1}{2}}$). Let the set of distinct eigenvalues of $\mathbf{A}$(the spectrum of $\mathbf{A}$) denote as $\sigma(\mathbf{A})$. The maximum magnitude of eigenvalues, denoted as $\rho(\mathbf{A}) = \max|\lambda|, \forall\lambda \in \sigma(\mathbf{A})$ has $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$ for every matrix norm (Meyer, 2000, p497), i.e., $|\lambda| \leq \|A\|$ for all $\lambda \in \sigma(\mathbf{A})$. Since all the components in the matrices $\mathbf{D}_{\mathbf{Es}}^{-\frac{1}{2}}, \mathbf{P}$, and $\mathbf{D}_{\pi} - \pi\pi'$ are regarding probabilities, the maximum absolute values of the components in the product for $\mathbf{A}$ is less than or equal to 1. Thus the maximum eigenvalue of the matrix $\mathbf{A}$ is equal to 1.

2.3.2 Asymptotic Distribution of $\tilde{Q}$

The statistical distance between the observed pseudocounts and their expectations (the external and fixed values) also represent the degree of model-data fit. If the distance is defined by $\tilde{Q} = (\bar{\mathbf{s}} - \mathbf{E}\bar{\mathbf{s}})\tilde{\mathbf{\Sigma}}^{-1}(\bar{\mathbf{s}} - \mathbf{E}\bar{\mathbf{s}})'$, and it is seen from theorem 2.3.1 and theorem 2.3.2 that the asymptotic distribution of the sequence $\sqrt{N}(\frac{\bar{s}}{N} - \tilde{\mathbf{P}}'\pi)$ for $\bar{s}$ is $N_Q(\mathbf{0}, \tilde{\mathbf{P}}'(\mathbf{D}_{\pi} - \pi\pi')\tilde{\mathbf{P}})$ with a nonsingular covariance matrix. As it is easy to see,

the expectation and covariance of $\bar{\mathbf{s}}$ is $E\bar{\mathbf{s}} = \tilde{\mathbf{P}}'\pi$ and $Var(\bar{\mathbf{s}}) = N\tilde{\mathbf{P}}'(\mathbf{D}_\pi - \pi\pi')\tilde{\mathbf{P}}$. Since $\frac{(\bar{\mathbf{s}}-\mathbf{E}\bar{\mathbf{s}})}{\sqrt{N}} = \sqrt{N}(\frac{\bar{\mathbf{s}}}{N} - \tilde{\mathbf{P}}'\pi)$, the following states the result for the asymptotic distribution of $\tilde{Q}$ (e.g., p163, Johnson and Wichern, 2002).

theorem 2.3.2 Asymptotic Distribution of $\tilde{Q}$ *The asymptotic distribution of item fit measure defined as* $\tilde{Q} = (\bar{\mathbf{s}} - \mathbf{E}\bar{\mathbf{s}})\tilde{\mathbf{\Sigma}}^{-1}(\bar{\mathbf{s}} - \mathbf{E}\bar{\mathbf{s}})'$ *is* χ^2 *with degree of freedom* Q *and the covariance matrix* $\tilde{\mathbf{\Sigma}}$ *is* $N\tilde{\mathbf{P}}'(\mathbf{D}_\pi - \pi\pi')\tilde{\mathbf{P}}$.

Let $\mathbf{\Sigma}$ denote the $2Q \times 2Q$ covariance matrix of the pseudocounts vector $\mathbf{s}$, i.e., $\mathbf{\Sigma} = \mathbf{N}\mathbf{P}'(\mathbf{D}_\pi - \pi\pi')\mathbf{P}$. Then the covariance matrix over the total pseudocounts vector $\bar{\mathbf{s}}$, $\tilde{\mathbf{\Sigma}}$, is $N\tilde{\mathbf{P}}'(\mathbf{D}_\pi - \pi\pi')\tilde{\mathbf{P}}$. The following section will introduce a consistent estimator of $\mathbf{\Sigma}$ and $\tilde{\mathbf{\Sigma}}$.

2.4 The Observed Covariance Matrix of Interrelations among Pseudocounts

Although the covariance matrix of pseudocounts $\mathbf{\Sigma} = \mathbf{P}'(\mathbf{D}_\pi - \pi\pi')\mathbf{P}$ has dimension of $2Q \times 2Q$ and the dimension of $\tilde{\mathbf{\Sigma}} = \tilde{\mathbf{P}}'(\mathbf{D}_\pi - \pi\pi')\tilde{\mathbf{P}}$ is $Q \times Q$, the estimation of $\mathbf{\Sigma}$ and $\tilde{\mathbf{\Sigma}}$ involves evaluating the $2T \times 2Q$ matrix $\mathbf{P}$, the $2T \times Q$ matrix of $\tilde{\mathbf{P}}$, and the $2T \times 2T$ matrix of $\mathbf{D}_\pi - \pi\pi'$. Note that T indicates all possible response patterns for the remaining J items. In dichotomous case, $T = 2^J$. For a long test, the numerical computation of $\mathbf{\Sigma}$ seems impractical for most operational work. To reduce the computation complexity, $\mathbf{\Sigma}$ is estimated from the observed covariance matrix $\mathbf{S}$ of interrelations among pseudocounts $\mathbf{s}$, and $\tilde{\mathbf{\Sigma}}$ is estimated from the observed covariance matrix $\tilde{\mathbf{S}}$ of interrelations among pseudocounts $\bar{\mathbf{s}}$. As is known that $\mathbf{n} \sim$

$\mathbf{M_{2T}(N,\pi)}$, a multinomial distribution with $2T-1$ parameters and covariance matrix $N(\mathbf{D}_\pi - \pi\pi')$. $\hat{P}_{tj}$ is a uniformly minimum variance unbiased estimator (UMVU) of $\pi_{tj}, \forall t = 1, 2, \cdots, T$, and $j = 0, 1$ (e.g., Lehmann and Casella, 1998, p106; Bickel and Docksum, 2001, p187). It is natural to think of the matrix $D_{\hat{\mathbf{P}}} - \hat{\mathbf{P}}\hat{\mathbf{P}}'$ as estimate of the matrix $\mathbf{D}_\pi - \pi\pi'$.

Let the vector $\mathbf{x_i}$ indicate the posterior contribution of the ith examinee on the studied item across the array of Q quadrature points given the response pattern $(U, Y = y_i), \forall i = 1, 2, \cdots, N$. Then $\mathbf{x_i}$ is a $2Q$ dimensional vector as $\mathbf{x_i} = (X^1_{i1}, X^1_{i2}, \cdots, X^1_{iQ}, X^0_{i1}, X^0_{i2}, \cdots, X^0_{iQ})$. The value of each component in the vector X_i is $X^j_{iq}, \forall q = 1, 2, \cdots, Q, i = 1, 2, \cdots, N$, and $j = 1, 0$. Or

$$\begin{aligned} X^1_{iq} &= UP(\theta = \theta_q | U = 1, Y = y_i), \\ X^0_{iq} &= (1-U)P(\theta = \theta_q | U = 0, Y = y_i). \end{aligned}$$

Therefore, a N by $2Q$ matrix representing the contributions of each examinee to the posteriors at Q quadrature points given the observed test data is available. If all of the item parameters are known constants, then each posterior can be thought of as a fixed value. And therefore, each component $X^j_{iq}, \forall q = 1, 2, \cdots, Q, i = 1, 2, \cdots, N$, and $j = 1, 0$ can be viewed as a random variable, because U is a Bernoulli random variable. Clearly, these $2Q$ random variables are not independent. The realization of each component X^j_{iq} constitutes a N by $2Q$ matrix, a much smaller and more manageable matrix for computation, which can then be used to estimate the covariance matrix among pseudocounts. It can be shown that the sum over all examinee's

posteriors (or a row vector in the N by $2Q$ matrix) is actually the pseudocounts vector $\mathbf{s}$. In this sense, the vector $\mathbf{x_i}$ can be regarded as the unit pseudocount, and for $\forall i = 1, 2, \cdots, N$, x_i can be viewed as the ith realization of the random vector $\mathbf{x} = (X_1^1, X_2^1, \cdots, X_Q^1, X_1^0, X_2^0, \cdots, X_Q^0)$. The N by $2Q$ matrix contains all information for each examinee' unit pseudocount on each quadrature point.

The observed covariance matrix of the vector $\mathbf{x}$ from sample data depicts the covariance $\mathbf{S}$ of interrelation of pseudocounts. The following section states the interrelations between the variance of the unit pseudocount X_q^j and overall pseudocounts $s_{qj}, \forall q = 1, 2, \cdots, Q$ and $j = 1, 0$. By definition, the sample variance of X_q^j is given by

$$Var(X_q^j) = \frac{1}{N}\sum_{i=1}^{N}\left(X_{iq}^j - \frac{s_{qj}}{N}\right)^2.$$

Denote $\bar{r}_{qj} = \frac{s_{qj}}{N} = \sum_{t=1}^{T} \hat{\mathbf{p}}_{tj} P_{tq}^j$, for j = 1, 0, then

$$\begin{aligned}
Var(X_q^j) &= \frac{1}{N}\sum_{i=1}^{N}\left(X_{iq}^j - \frac{s_{qj}}{N}\right)^2 \\
&= \sum_{t=1}^{T}\frac{n_{tj}}{N}(P_{tq}^j - \frac{s_{qj}}{N})^2 + \frac{N - \sum_{t=1}^{T} n_{tj}}{N}\frac{s_{qj}^2}{N^2} \\
&= \sum_{t=1}^{T}\hat{\mathbf{p}}_{tj}{P_{tq}^j}^2 - 2\sum_{t=1}^{T}\hat{P}_{tj}P_{tq}^j\bar{r}_{qj} + \bar{r}_{qj}^2 \\
&= \sum_{t=1}^{T}\hat{P}_{tj}{P_{tq}^j}^2 - \left(\sum_{t=1}^{T}\hat{P}_{tj}P_{tq}^j\right)^2 \\
&= {P_q^j}^T(\mathbf{D}_{\hat{p}} - \hat{\mathbf{p}}\hat{\mathbf{p}}')P_q^j.
\end{aligned}$$

When the item parameters are all known constants, the posterior P_{tq}^1 is also known and fixed, $\forall t = 1, 2, \cdots, T$, and $q = 1, 2, \cdots, Q$. The difference between the variance for the unit pseudocount $Var(X_q^1)$ and the average sample variance of the

total pseudocounts $\frac{1}{N}Var(s_{q1})$ is $Var(X_q^1) - \frac{1}{N}Var(s_{q1}) = \sum_{t=1}^{T}(\hat{P}_{t1} - \pi_{t1})P_{tq}^{1\,2} - \left(\sum_{t=1}^{T}(\hat{P}_{t1} - \pi_{t1})P_{tq}^1\right)\left(\sum_{t=1}^{T}(\hat{P}_{t1} + \pi_{t1})P_{tq}^1\right)$. Since the sample proportion $\hat{P}_{t1}$ is a consistent estimator for the parameter π_{t1}, $\hat{P}_{t1} \rightarrow \pi_{ti}$ in probability as $N \rightarrow \infty$. Thus, as the sample size N goes to infinity, $Var(X_q^1) \rightarrow \frac{1}{N}Var(s_{q1})$ in probability. That is, $Var(X_q^1)$ is consistent for estimating $\frac{1}{N}Var(s_{q1}), \forall q = 1, 2, \cdots, Q$. Similarly, it can be shown that the $Var(X_q^0)$ is a consistent estimator of $\frac{1}{N}Var(S_{q0}), \forall q = 1, 2, \cdots, Q$.

Now consider the relations between the covariance $Cov(X_q^j, X_{q'}^{j'})$ and the covariance $Cov(s_{qj}, s_{q'j'}), \forall q, q' = 1, 2, \cdots, Q$, and $j, j' = 1, 0$. First express $Cov(s_{qj}, s_{q'j'})$ as

$$\begin{aligned} Cov(s_{qj}, s_{q'j'}) &= Cov(\sum_{t=1}^{T} n_{tj}P_{tq}^j, \sum_{t=1}^{T} n_{tj'}P_{tq'}^{j'}) \\ &= Cov(P_q^{j\,T}\mathbf{n}, \mathbf{n}'\mathbf{P}_{\mathbf{q}'}^{\mathbf{j}'}) \\ &= NP_q^{j\,T}(\mathbf{D}_\pi - \pi\pi')P_{q'}^{j'}. \end{aligned}$$

The vectors P_q^j and $P_{q'}^{j'}$ are two columns in the matrix $\mathbf{P}$ with row q, q' and column j, j', respectively.

Next, study the covariance $Cov(X_q^j, X_{q'}^{j'})$. By definition,

$$\begin{aligned} Cov(X_q^j, X_{q'}^{j'}) &= \frac{1}{N}\sum_{i=1}^{N}(X_{iq}^j - \frac{s_{qj}}{N})(X_{iq'}^{j'} - \frac{s_{q'j'}}{N}) \\ &= \frac{1}{N}\sum_{i=1}^{N}(X_{iq}^j X_{i'q'}^{j'} - X_{iq}^j\bar{r}_{q'j'} - X_{iq'}^{j'}\bar{r}_{qj}) + \bar{r}_{qj}\bar{r}_{q'j'} \\ &= P_q^{j\,T}\mathbf{D}_\pi P_{q'}^{j'} - P_q^{j\,T}\hat{\mathbf{p}}\hat{\mathbf{p}}'P_{q'}^{j'} \\ &= P_q^{j\,T}(\mathbf{D}_{\hat{\mathbf{p}}} - \hat{\mathbf{p}}\hat{\mathbf{p}}')P_{q'}^{j'}. \end{aligned}$$

Again, it is seen that $Cov(X_q^j, X_{q'}^{j'}) \rightarrow \frac{1}{N}Cov(s_{qj}, s_{q'j'})$. It is not hard to find that

$Cov(\bar{s}_q, \bar{s}'_q) = N\tilde{P}_q^T(\mathbf{D}_\pi - \pi\pi')\tilde{\mathbf{P}}_{\mathbf{q}'}$, $\forall q, q' = 1, 2, \cdots, Q$. Form the $N \times Q$ posterior matrix $\tilde{\mathbf{X}}$ with each row vector $\tilde{\mathbf{x}}_i, \forall i = 1, 2, \cdots, N$ representing the ith examinee's and number of Q posteriors, then $\tilde{\mathbf{x}}_i = (X_{i1}, X_{i2}, \cdots, X_{iQ}), \forall i = 1, 2, \cdots, N$, where

$$X_{iq} = ((P_{i1}^1)^U(P_{i1}^0)^{1-U}, (P_{i2}^1)^U(P_{i2}^0)^{1-U}, \cdots, (P_{iQ}^1)^U(P_{iQ}^0)^{1-U}).$$

In the same way, the vector $\tilde{\mathbf{x}}_\mathbf{i}$ is one realization of the vector $\tilde{\mathbf{x}} = (X_1, X_2, \cdots, X_Q)$. Let the covariance of the vector $\tilde{\mathbf{x}}$ denote $\tilde{\mathbf{S}} = Cov(\tilde{\mathbf{x}})$. It can be seen that $Cov(X_q, X_{q'}) = \tilde{P}_q^T(\mathbf{D}_{\hat{\mathbf{p}}} - \hat{\mathbf{p}}\hat{\mathbf{p}}')\tilde{P}_{q'}$

In a summary, the observed covariance matrix $\mathbf{S}$ of the interrelation among pseudocounts is a consistent estimator of the average covariance of pseudocounts vector $\mathbf{s}$. $\mathbf{S}$ can be arbitrarily close to $\frac{1}{N}\mathbf{\Sigma}$ when the sample size N is large enough. Similarly, the observed matrix $\tilde{\mathbf{S}}$ is a consistent estimator of $\frac{1}{N}\tilde{\mathbf{\Sigma}}$. The noticeable computational simplicity can be obtained using $\mathbf{S}$, which is a constructed N by $2Q$ matrix of posteriors. The simplification of the computational complexity for the covariance matrix among pseudocounts make the hypothesis testing of goodness of fit at the item level feasible using the measure of item fit Q_{DM}^*.

2.5 Estimation of the Asymptotic Distribution for Q_{DM}^*

The true asymptotic distribution of Q_{DM}^* is a function of the covariance matrix $\mathbf{\Sigma}$ of pseudocounts. The relations of the asymptotic distribution with the covariance $\mathbf{\Sigma}$ rely on the non-zero eigenvalues λ's from the matrix $\mathbf{\Sigma}$. The asymptotic distribution of Q_{DM}^* can be written as $\sum_{i=1}^m \lambda_i \chi_{i(1)}^2$ and the nonzero coefficients λ's comes from

the matrices $\boldsymbol{\Sigma}$. Denote the non-zero eigenvalues from the observed covariance S as $\hat{\lambda}$'s. Then the differences between the true asymptotic distribution and the estimated asymptotic distribution is $\sum_{i=1}^{m}(\hat{\lambda}_i - \lambda_i)\chi^2_{i(1)}$. It is easy to see that the estimated distribution is arbitrarily close to the true asymptotical distribution as long as the $\hat{\lambda}'s$ are arbitrarily close to the true λ's. Obviously, as $N \rightarrow \infty$, due to the consistency of $\hat{\boldsymbol{\Sigma}}$ to $\boldsymbol{\Sigma}$, $\hat{\lambda}_i$ is arbitraily close to λ_i, $\forall i = 1, 2, \cdots, m$.

For the estimate of the asymptotic distribution of $\tilde{Q}$, replace the covariance matrix $\tilde{\boldsymbol{\Sigma}}$ in the middle of $(\bar{\mathbf{s}} - \mathbf{E}\bar{\mathbf{s}})\tilde{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{s}} - \mathbf{E}\bar{\mathbf{s}})'$ with its consistent estimator. Then asymptotic distribution of the estimate is arbitrarily close to its true asymptotic distribution. Therefore, the asymptotic distribution of the fit measures Q^*_{DM} and $\tilde{Q}$ with true covariance matrix among pseudocounts are the same as the asymptotic distributions of fit measures Q^*_{DM} and $\tilde{Q}$, respectively, with observed covariance matrix of interrelations among pseudocounts as their corresponding consistent estimators of the true covariance matrix.

Assuming item parameters known constants is not realistic in many applications. This section will investigate the relations between the item parameter estimates and asymptotic distribution of the reformulated item fit measure Q^*_{DM} for data-based item parameter estimates.

Since item response function f_{qj} are continuous function of item parameters given each quadrature point $\theta_q, \forall q = 1, 2, \cdots, Q$ and $j = 0, 1$, $f_{qj}(\hat{a}_n, \hat{b}_n, \hat{c}_n, \theta_q) \rightarrow f_{qj}(a, b, c, \theta_q)$ in probability as $n \rightarrow \infty$, in short, $\hat{f}_{qj} \rightarrow f_{qj}$, if both item and ability parameters are consistent estimates (e.g., p124,Rao, 1976; p74, 8.4, Lehmann, and Casella, 1998). It

is also not hard to demonstrate that $\hat{l}_{tq} \rightarrow l_{tq}$ in probability, $\hat{\pi}_{tj} \rightarrow \pi_{tj}$ in probability, $\hat{s}_{qj} \rightarrow s_{qj}$ in probability, and $\hat{E}s_{qj} \rightarrow Es_{qj}$ in probability, $\forall q = 1, 2, \cdots, Q, j = 0, 1$, and $\forall t = 1, 2, \cdots, T$. In the same way, the estimates of Q^*_{DM} and $\tilde{Q}$ tend to true Q^*_{DM} and $\tilde{Q}$, respectively in probability. Moreover, by convergence together theorem (e.g., p122, Rao, 1976; p91, Durret, 1996), the estimates of s_{qj}, Q^*_{DM}, and $\tilde{Q}$ have the same asymptotic distribution as those of s_{qj}, Q^*_{DM}, and $\tilde{Q}$, correspondingly, $\forall q = 1, 2, \cdots, Q, j = 0, 1$. Therefore, suppose the consistent estimates of item parameters are available, the results on the item fit measure Q^*_{DM} and its corresponding asymptotic distribution can be extended to the situation in which item parameters are data-based estimates in theory.

Chapter 3

Simulation Studies on Item Fit

Several simulation studies on the item fit measure Q^*_{DM} are presented in this chapter. One important purpose for the simulation studies is to examine how large the additional errors might be induced by the approximation for the asymptotic distribution based on the observed covariance compared to the true asymptotic distribution, and to find out what conditions can make the approximation practically useful. To investigate the accuracy of the approximation, a test consisting of 15 items is simulated. Such a short test is chosen because most personal computers can handle the computation involving all possible response patterns of 15 items, which is required for computing the true asymptotic probability. For dichotomously scored responses, there are $2^{15} = 32678$ possible response patterns in all. Thus, the true asymptotic distribution, the approximation of the true asymptotic distribution based on the observed covariance matrix of interrelations among pseudocounts, and the approximation on the basis of data-based item parameter estimates can be compared to each other. For a longer test (e.g., a 30-item test), the possible response patterns may be too huge (e.g., 1073741824 for a 30-item test) to compute the true asymptotic probabilities.

Without the true asymptotic probabilities, it is difficult to have an intuitive sense of how good is the approximation. The comparison of the true parameters and parameter estimates (e.g., the true covariance among pseudocounts versus the covariance estimate, the true asymptotic probability versus the approximation, the true eigenvalues versus the estimated eigenvalues from the observed covariance matrix, the true item parameters versus the item parameter estimates) is viewed as an oracle analysis. In applications, there is no need to compute all possible response patterns for the sake of the true covariance matrix among pseudocounts, if the approximation is sufficiently close to the true value or the induced errors are negligible for practical use. To compute the true covariance matrix among pseudocounts here and the true asymptotic distribution for a given Q^*_{DM} is merely for the convenience of the comparison to which one can see how good the approximation can be. According to this asymptotic method and approximation approach, there should be no practical concerns on the computation of item fit analysis for longer tests. Therefore, the method is not limited to short tests only. It can be applied to longer tests as long as the sample size is large enough so that the approximation work well.

Three different sample sizes are chosen for this study to determine how large the sample sizes are sufficient for this asymptotic method, and attempt to provide a guideline on how large sample size is sufficient for the method to work well. The 15 item parameters are also generated from computers. Discriminating power parameters are simulated from uniform distribution ranging from .6 $\sim$ 2.6, i.e., $U(.6, 2.6)$, difficulty parameters are generated from standard normal distribution $N(0, 1)$, and

Table 3.1: True Item Parameters for the Test of 15 Items

Item	Discrimination a	Difficulty b	Asymptote c
1	.672	1.410	.177
2	1.652	1.493	.013
3	.747	.935	.005
4	1.486	1.706	.165
5	1.286	.967	.080
6	1.357	.820	.086
7	1.140	-.411	.159
8	1.107	1.060	.083
9	1.465	.388	.085
10	.920	1.643	.145
11	.740	-.668	.173
12	.803	1.125	.040
13	1.407	-.451	.067
14	.662	.077	.124
15	1.845	1.166	.148

the asymptote parameters are from uniform $U(0, .25)$. Table 3.1 contains all of the true parameter values for the 15 items.

Three groups of examinees are generated from $N(0, 1)$ with sample sizes 500, 1000, and 5000, which represent small, medium, and large samples, respectively. For each sample, dichotomous response data are simulated from 3PL IRT models. To account for the randomness from the response data, replications (1000) for each sample size will be conducted. More specifically, the 15-item test will be administrated to 1000 groups of examinees with sample size 500 each from $N(0, 1)$, and 1000 with sample size 1000, and 1000 with sample size 5000. Combined with the sample size and replication conditions, there are in all 3000 data sets yielded for the simulation studies.

3.1 Type I Error Rates

To allow comparisons, the true asymptotic distribution,the approximation of the true asymptotic distribution based on the observed covariance matrix of interrelations among pseudocounts, and the estimated asymptotic distribution on the basis of item parameter estimates as well are computed alone with the corresponding item fit measure Q^*_{DM}. Type I error rates are calculated and compared across different sample sizes (e.g., 500, 1000, and 5000). Under the null hypothesis that the simulated response data from the 3PL model fit the hypothetical 3PL model (in this example, the same form of mathematic model is assumed for all items in the short test—3PL model), the observed item fit measure Q^*_{DM} is asymptotically distributed as a quadratic form of normal variables, which is addressed in Chapter 2. For a given observed item fit statistic Q^*_{DM}, the asymptotic probability of observing such a value or greater can be evaluated through the routine by Davies (1980). For each item and each replication, count the number of times for the hypothetic item model being rejected. If the number is greater than 50 over 1000 replications (i.e., the type I error rate is greater than .05), it is said the type I error is greater than what is expected. Otherwise, the type I error rate would be acceptable. Table 3.2-3.4 shows type I error rates for each item in the test over 1000 replications across three different sample sizes (e.g., 500, 1000, and 5000).

A good model-data fit test requires low type I error rate. The lower the type I error rate, the less mistakes that would be made when to accept a correct hypothesis.

Table 3.2: Type I Error Rate for Sample Size 500

Item	Type I Error			RMSE		
	True(Full)	True(Appr.)	Item Esti.	a	b	c
1	.026	.023	.000	.187	.254	.040
2	.029	.029	.006	.401	.190	.019
3	.016	.017	.002	.254	.257	.089
4	.020	.023	.001	.506	.276	.023
5	.011	.013	.000	.228	.153	.030
6	.016	.019	.000	.234	.135	.030
7	.023	.024	.002	.169	.093	.035
8	.018	.020	.001	.200	.178	.033
9	.011	.014	.001	.231	.108	.040
10	.026	.031	.000	.221	.247	.026
11	.023	.025	.003	.115	.136	.036
12	.017	.013	.000	.233	.224	.062
13	.017	.018	.008	.630	.112	.092
14	.031	.035	.001	.148	.243	.081
15	.019	.020	.000	1.148	.152	.023

Table 3.3: Type I Error Rate for Sample Size 1000

Item	Type I Error			RMSE		
	True(Full)	True(Appr.)	Item Esti.	a	b	c
1	.027	.023	.000	.162	.174	.035
2	.012	.012	.000	.292	.092	.012
3	.012	.012	.000	.208	.180	.070
4	.014	.012	.000	.303	.146	.018
5	.016	.013	.000	.192	.098	.022
6	.019	.017	.000	.203	.092	.022
7	.024	.020	.003	.123	.082	.032
8	.010	.010	.000	.184	.105	.025
9	.011	.010	.000	.179	.089	.030
10	.033	.029	.000	.199	.148	.022
11	.020	.020	.001	.087	.122	.037
12	.020	.016	.000	.192	.146	.046
13	.018	.013	.003	.193	.113	.075
14	.025	.023	.000	.118	.211	.070
15	.020	.019	.000	.317	.085	.018

Table 3.4: Type I Error Rate for Sample Size 5000

Item	Type I Error			RMSE		
	True(Full)	True(Appr.)	Item Esti.	a	b	c
1	.054	.045	.000	.079	.083	.023
2	.020	.019	.000	.130	.043	.004
3	.049	.045	.001	.099	.093	.035
4	.040	.034	.000	.175	.061	.008
5	.039	.036	.000	.093	.046	.011
6	.037	.032	.000	.096	.042	.011
7	.036	.033	.000	.070	.062	.035
8	.045	.038	.000	.084	.051	.012
9	.029	.026	.000	.090	.042	.014
10	.050	.040	.000	.099	.072	.013
11	.047	.043	.000	.047	.097	.037
12	.035	.032	.000	.077	.065	.019
13	.020	.020	.001	.095	.067	.034
14	.044	.039	.000	.056	.114	.038
15	.026	.024	.000	.176	.042	.009

To examine the type error rates for item fit test, the data are generated from the particular mathematic models (e.g., the 3PL model) and fit back into the same item model—an obvious known fact or correct hypothesis. Therefore, the item fit test, if it is right, should provide useful information to accept the correct hypothesis except some acceptable level of errors (Type I error) due to randomness; or the item fit test is simply employed to verify the known fact. The type I error rates in the tables are calculated based on 1000 replications for each sample size. Table 3.2 through 3.4 give the type I error rates when item parameters are known (denoted as "Full" and "Appr.") and type I error error rates when item parameters are estimated from the response data (denoted as "Item Esti.") along with the root mean square errors (denoted as "RMSE") for each item parameter estimates.

It can be seen from the three tables (table 3.2, 3.3, and 3.4) that the type I error rates across different sample sizes are basically very low, lower than .05, the level of significance. Only one item (the first item in the 5000 case in table 3.4) has type error rate .054, a little bit bigger than the significant level, on the true asymptotic distribution.

One major feature of the type I error rates in the tables is when item parameters are known constants, the type I error rate based on the true asymptotic distribution is close to their counterpart from the approximation by the observed covariance matrix. However, the type I error rates from the data-based item parameter estimates are in general less than those from the true item parameters and are very conservative regardless of the sample sizes. It can be seen from these tables that most of the items the type I error rates are near to zero.

As seen in any estimation programs in IRT, item parameter estimates contain estimation errors even if the data adequately fit the mathematical models used for the estimation. To examine the conservative performance of Q^*_{DM} under the circumstances of the item parameter estimates, root mean square errors (RMSE) of the item parameter estimates are calculated from the data sets. RMSE is defined as the square root of the mean squared difference between the item parameter estimates and the true item parameters over r replications (r in this example is 1000). Let η denote as the item parameter (e.g., discriminating power parameter a, or difficulty parameter b, or asymptote parameter c) and $\hat{\eta}$ as the item parameter estimates. Then RMSE

can be calculated by

$$RMSE(\eta) = \sqrt{\frac{\sum_{i=1}^{r}(\hat{\eta}_i - \eta_i)^2}{r}}.$$

RMSE provides a summary index of assessing the accuracy of item parameter estimates. Apparently, the larger RMSE of the item parameter estimates, the worse of the estimation. For a simulation study, an adequate fit of model and data is assumed, and thus the difference in the item parameter estimates may depend on the estimation procedures and some other factors (e.g., sample size of examinees).

Table 3.2 through 3.4 also contain the RMSE over 1000 replications for each item parameter in the test. The estimation procedure used in this study for BILOG-MG3 is Bayesian MML with default item prior distributions. That is, for all item parameters, $a \sim lognormal(0, 0.5), b \sim N(0, 2), c \sim beta(5, 17)$. It shows from these three tables that the RMSE decreases as the sample size increases, indicting that better item parameter estimates are obtained, which is expected. In general, the RMSE for the sample size equal to 500 is the largest and for 5000 the RMSE is the smallest. For the same sample size, the RMSE for discriminating power parameter is in general larger than that of difficulty and asymptote parameters.

3.2 Coefficients for the Asymptotic Distributions

Since the asymptotic distribution depends on the coefficients in the linear combination, i.e., the eigenvalues extracted from the covariance among pseudocounts, it is important to compare the coefficients from the true covariance matrix (the full covariance matrix that comes from evaluating all possible response patterns for a

given test), approximation of the covariance matrix, and estimated covariance matrix on the data-based item parameters estimates across different sample sizes. The purpose of comparing those coefficients is to examine how much additional error is induced through the coefficients of the asymptotic distribution. Table 3.5 through table 3.11 include 20 ordered positive eigenvalues extracted from the true covariance matrix (i.e., Table 3.5) and from the approximated covariance matrix of the observed covariance among pseudocounts as well (Table 3.6 through table 3.8 show the 20 ordered positive eigenvalues from the true item parameters; table 3.9 through table 3.11 from the data-based item parameter estimates). In these tables, the rows represent the 20 positive eigenvalues and the columns indicate the 15 items in the test. The other extracted eigenvalues are omitted and not used for calculating the asymptotic probabilities due to their trivial magnitudes. Note the values in these tables are from one replication. Similar results can be obtained from other 999 replications and hence are not reported here. The 20 ordered positive eigenvalues from the true covariance matrix, which depends only on the number of items, are used to compute true asymptotic distributions; the 20 ordered positive eigenvalues extracted from the observed covariance matrix of interrelations among pseudocounts from true item parameters are used to compute the approximation of the asymptotic probabilities; the 20 ordered positive eigenvalues from the observed covariance matrix of interrelations among pseudocunts based on the item parameter estimates are the coefficients for computing the estimated asymptotic probabilities.

It can be shown from these tables that for each sample size the 20 eigenvalues

Table 3.5: The 20 Positive Eigenvalues from True Covariance Matrix

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	.01	0	.01	.01	0	.01	.01	.01	.01	.01	0	.01	0	0
.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
.02	.02	.01	.02	.02	.01	.01	.02	.02	.02	.02	.01	.01	.01	.01
.03	.03	.03	.03	.02	.02	.03	.03	.03	.03	.03	.02	.02	.03	.02
.04	.05	.03	.04	.05	.02	.03	.04	.04	.04	.04	.03	.03	.03	.03
.06	.07	.07	.07	.06	.04	.07	.06	.07	.07	.07	.06	.05	.06	.06
.09	.1	.07	.08	.11	.04	.08	.08	.09	.1	.09	.06	.07	.06	.06
.13	.13	.13	.14	.14	.1	.14	.13	.13	.14	.14	.13	.11	.13	.13
.18	.2	.15	.18	.23	.1	.17	.17	.18	.2	.19	.13	.16	.14	.13
.23	.24	.24	.25	.26	.19	.25	.23	.24	.25	.25	.24	.2	.23	.24
.33	.34	.28	.32	.39	.2	.32	.31	.32	.36	.33	.26	.29	.27	.26
.37	.38	.39	.4	.44	.34	.4	.38	.39	.42	.4	.39	.33	.37	.39
.52	.53	.47	.53	.59	.37	.53	.5	.51	.56	.53	.44	.47	.45	.44
.55	.57	.58	.58	.65	.56	.59	.56	.56	.62	.59	.58	.52	.55	.58
.77	.78	.71	.79	.84	.63	.8	.75	.76	.82	.79	.69	.72	.67	.71
.8	.82	.79	.83	.87	.82	.85	.79	.79	.86	.82	.78	.78	.79	.78
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 3.6: 20 Eigenvalues for True Item Parameters ($N = 500$)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	0	0	.01	.01	0	0	0	0	0	0	0	0	.01	0
.01	0	0	.01	.01	.01	0	.01	.01	.01	0	.01	0	.01	0
.02	.01	.01	.02	.01	.01	0	.01	.01	.01	.01	.01	0	.01	.01
.04	.01	.01	.03	.02	.02	0	.03	.03	.03	.02	.01	.01	.02	.01
.04	.03	.03	.04	.04	.03	.01	.03	.03	.03	.03	.04	.01	.04	.04
.1	.05	.05	.06	.06	.05	.02	.08	.07	.08	.05	.04	.02	.06	.04
.1	.09	.09	.11	.1	.09	.03	.08	.09	.08	.09	.09	.04	.1	.09
.19	.12	.12	.14	.14	.14	.06	.19	.15	.18	.13	.1	.06	.13	.12
.22	.2	.2	.21	.21	.2	.08	.19	.19	.2	.18	.2	.12	.21	.19
.38	.25	.23	.29	.27	.32	.18	.35	.27	.32	.28	.21	.14	.31	.27
.42	.39	.37	.41	.39	.36	.21	.37	.4	.35	.35	.36	.27	.4	.38
.59	.44	.42	.51	.52	.57	.34	.56	.47	.55	.56	.43	.41	.58	.53
.7	.52	.62	.64	.61	.59	.55	.59	.63	.55	.59	.64	.5	.61	.59
.76	.72	.68	.74	.74	.74	.63	.77	.72	.8	.75	.65	.62	.68	.8
.93	.82	.85	.88	.77	.89	.78	.8	1.02	.82	.99	.87	.74	.85	.82
1.07	1.06	1.02	1.02	.99	1	.99	1.01	2.73	.95	1.45	1.1	1.02	1.01	1.03

Table 3.7: 20 Eigenvalues for True Item Parameters ($N = 1000$)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	0	0	.01	.01	0	0	.01	0	.01	0	0	0	.01	.01
.02	.01	.01	.01	.01	.01	0	.01	.01	.01	0	.01	0	.01	.01
.02	.01	.01	.01	.02	.01	0	.01	.01	.02	.01	.01	0	.02	.01
.04	.02	.03	.02	.03	.02	0	.03	.03	.03	.01	.02	.01	.02	.03
.04	.03	.03	.04	.04	.03	.01	.04	.03	.04	.04	.04	.01	.04	.04
.09	.05	.08	.06	.07	.05	.01	.08	.07	.08	.04	.04	.03	.05	.06
.1	.09	.08	.09	.09	.08	.03	.09	.09	.09	.09	.1	.04	.09	.09
.19	.13	.17	.13	.16	.13	.05	.18	.16	.18	.09	.11	.07	.13	.15
.21	.21	.17	.18	.2	.17	.08	.19	.18	.2	.18	.21	.13	.2	.18
.37	.28	.32	.28	.35	.28	.18	.35	.28	.36	.24	.24	.15	.29	.31
.41	.4	.35	.37	.38	.3	.2	.4	.31	.36	.35	.39	.28	.38	.37
.6	.49	.54	.51	.58	.5	.34	.55	.47	.57	.51	.46	.32	.47	.5
.64	.58	.54	.63	.64	.55	.55	.66	.57	.59	.57	.66	.51	.6	.56
.8	.75	.73	.78	.79	.71	.55	.79	.73	.78	.76	.67	.51	.65	.75
.88	.84	.77	.86	.86	.8	.8	.91	1	.86	.8	.91	.78	.85	.85
1.03	1.02	.97	.97	1.02	.95	.99	1.01	1.59	.99	1	1.06	1.01	1	.99

from the estimated covariance matrix across three different sample sizes are very close to their counterparts from the true covariance matrix no matter whether the item parameters are true constants or data-based estimates. These values are the estimated coefficients for the linear combination of χ_1^2 random variables, which are eventually used to calculate the asymptotic probabilities. Except for the 20 values from the true covariance matrix in Table 3.5, the coefficients in Table 3.6 through Table 3.11, which are from observed covariance matrices of pseudocounts, are data-based estimates and vary as data change. And so do the resulting approximation of true asymptotic probabilities. For example, over 1000 replications of the 15-item test with 500 examinees, there are 1000 different observed covariance matrices of

Table 3.8: 20 Eigenvalues for True Item Parameters ($N = 5000$)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	0	0	.01	.01	0	0	.01	0	.01	0	.01	0	.01	0
.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	0	.01	0	.01	.01
.02	.01	.01	.02	.02	.01	.01	.02	.01	.01	.01	.02	0	.02	.01
.04	.02	.03	.02	.02	.02	.02	.03	.03	.03	.01	.03	.01	.03	.02
.04	.03	.03	.04	.04	.04	.03	.04	.04	.04	.04	.04	.01	.04	.04
.09	.06	.08	.06	.06	.06	.06	.08	.06	.08	.04	.08	.03	.07	.06
.1	.08	.08	.09	.09	.09	.08	.09	.08	.09	.09	.1	.05	.09	.09
.18	.13	.18	.14	.14	.13	.15	.19	.14	.18	.1	.17	.07	.16	.13
.21	.18	.18	.2	.2	.19	.18	.2	.18	.2	.2	.2	.13	.2	.19
.34	.28	.34	.27	.29	.25	.29	.37	.29	.34	.23	.32	.16	.31	.25
.4	.35	.35	.39	.38	.37	.36	.38	.36	.38	.37	.39	.31	.38	.37
.55	.52	.55	.51	.53	.47	.52	.58	.53	.56	.41	.56	.33	.54	.44
.61	.58	.56	.61	.62	.57	.56	.61	.55	.59	.58	.61	.52	.59	.57
.8	.76	.76	.76	.74	.68	.75	.8	.66	.8	.63	.78	.57	.73	.72
.82	.78	.8	.88	.85	.77	.8	.86	.77	.83	.82	.83	.78	.83	.83
1	.95	.99	1	1.01	.99	1.01	.99	1	.97	1	1.01	1.02	1	.98

pseudocounts, and correspondingly the 20 ordered positive eigenvalues extracted from these matrices vary across data sets. However, it is found that for the 20 ordered positive eigenvalues extracted from each observed covariance matrix of pseudocounts, the differences between their true counterparts are so small that the approximation of the distribution is close to the true asymptotic distribution even for small sample size of 500. Similar results are also found for the case of the sample size 1000 and 5000. In addition, as the sample size increases, the the observed covariance matrix of pseudocounts become closer to the true covariance matrix of pseudocounts, and hence the approximation of the asymptotic distribution gets closer to its true asymptotic distribution.

Table 3.9: 20 Eigenvalues for Item Parameter Estimates ($N = 500$)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	0	0	0	0	0	0	0	0	.01	0	0	0	.01	0
.01	0	0	.01	.01	0	0	.01	.01	.01	0	0	0	.01	.01
.01	.01	.01	.01	.01	.01	0	.01	.01	.01	.01	.01	0	.01	.01
.04	.02	.01	.03	.02	.02	0	.03	.03	.02	.01	.02	0	.03	.02
.04	.02	.04	.04	.04	.04	.01	.03	.03	.03	.03	.03	.01	.04	.04
.09	.05	.04	.07	.06	.05	.01	.07	.07	.07	.05	.04	.02	.07	.05
.1	.08	.09	.1	.1	.1	.03	.09	.09	.1	.09	.08	.02	.1	.09
.18	.13	.12	.15	.15	.13	.05	.17	.16	.16	.15	.11	.06	.16	.12
.23	.16	.19	.21	.21	.21	.08	.2	.2	.22	.19	.19	.09	.21	.21
.4	.27	.28	.3	.3	.28	.19	.36	.29	.31	.31	.25	.15	.32	.26
.42	.28	.37	.4	.39	.38	.2	.37	.39	.38	.37	.33	.29	.41	.39
.62	.45	.49	.53	.56	.54	.35	.58	.49	.54	.58	.56	.35	.62	.51
.63	.47	.61	.64	.62	.58	.58	.59	.63	.63	.62	.59	.52	.64	.64
.84	.7	.73	.77	.77	.79	.64	.77	.76	.78	.81	.78	.8	.86	.76
.86	.76	.82	.93	.79	.79	.83	.82	1	.92	1	.8	1	.92	.87
1.01	1.01	1	1.01	1	1	1.01	1	2.99	1.01	1.68	1	1.25	1	1

3.3 Item Misfit and Power with Known Item Parameters

A good significance test also requires higher power for detecting model-data misfit. The higher the power for a hypothesis test, the higher the probability to reject the null hypothesis when it is actually incorrect. In this section, power is not computed analytically for the hypothesis testing, but is estimated empirically through simulated data. To estimate the power, for instance, the 3PL model is used to generate dichotomous response data, then fit the data generated by the 3PL model with the 2PL or the 1PL models, respectively. The Type I error rate is expected be low when fitting the data back with the 3PL model, but the power is expected high when fitting with

Table 3.10: 20 Eigenvalues for Item Parameter Estimates ($N = 1000$)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	0	.01	.01	.01	0	0	.01	0	.01	0	.01	0	.01	.01
.02	.01	.01	.01	.01	0	0	.01	.01	.01	0	.01	0	.01	.01
.02	.01	.01	.02	.02	.01	0	.01	.01	.02	.01	.02	0	.02	.01
.04	.02	.03	.02	.03	.02	0	.03	.03	.03	.01	.02	.01	.03	.03
.04	.03	.04	.04	.04	.04	.01	.04	.03	.04	.04	.04	.01	.04	.04
.1	.06	.08	.06	.08	.05	.01	.08	.08	.08	.04	.04	.03	.07	.06
.1	.09	.09	.09	.09	.09	.04	.09	.09	.1	.09	.09	.04	.1	.09
.19	.14	.18	.13	.17	.12	.05	.18	.17	.18	.12	.12	.07	.16	.15
.21	.2	.18	.19	.19	.19	.09	.19	.21	.2	.19	.2	.13	.2	.19
.38	.29	.35	.28	.36	.26	.18	.36	.31	.37	.27	.26	.16	.32	.31
.42	.32	.37	.39	.38	.36	.2	.4	.37	.38	.37	.36	.31	.39	.39
.6	.49	.58	.52	.6	.52	.35	.56	.5	.6	.57	.54	.37	.6	.5
.64	.51	.62	.65	.64	.58	.57	.66	.58	.62	.58	.61	.54	.6	.59
.84	.74	.8	.79	.81	.75	.59	.82	.77	.8	.83	.77	.56	.74	.74
.87	.77	.82	.91	.82	.8	.83	.86	1	.89	.99	.84	.82	.86	.88
1	1.01	1	1	1	1	1.01	1	2.18	1	1.18	1	1	1	1

the data with the 2PL or 1PL models over 1000 replications. Similarly, low type I error rates are expected when fitting the dichotomous response data generated by the 2PL model with the hypothetical 2PL model over 1000 replications, whereas power is expected high when fitting with the data with the hypothetical 1PL model. Table 3.12 through 3.14 show the power for all items at nominal level in the test for different sample sizes provided all item parameters are known constants.

From table 3.12, it can be seen easily that fitting the data generated by the 3PL model with the hypothetical 2PL or 1PL model is not adequate given the item parameters are known. Most times over 1000 replications the incorrect hypothesis is rejected, which makes the correction decision on the model-data misfit tests. From

Table 3.11: 20 Eigenvalues for Item Parameter Estimates ($N = 5000$)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
.01	0	0	.01	.01	.01	0	.01	0	.01	0	.01	0	.01	0
.01	.01	.01	.01	.01	.01	.01	.01	.01	.01	0	.01	0	.01	.01
.02	.01	.01	.02	.02	.01	.01	.02	.01	.02	.01	.02	0	.02	.01
.04	.02	.03	.02	.03	.02	.03	.03	.03	.03	.01	.03	.01	.03	.02
.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04	.01	.04	.04
.09	.06	.08	.06	.06	.07	.07	.08	.06	.08	.04	.08	.03	.08	.06
.1	.09	.09	.09	.09	.09	.09	.1	.09	.1	.09	.1	.04	.09	.09
.19	.14	.18	.14	.14	.14	.17	.18	.14	.18	.11	.17	.08	.17	.13
.22	.19	.19	.2	.2	.2	.19	.21	.19	.21	.2	.2	.13	.2	.2
.35	.28	.35	.28	.3	.26	.34	.35	.29	.34	.25	.32	.16	.33	.27
.41	.37	.36	.39	.39	.38	.37	.39	.37	.4	.38	.39	.33	.39	.39
.56	.52	.57	.53	.55	.5	.54	.57	.52	.56	.48	.58	.33	.57	.46
.62	.6	.58	.62	.63	.6	.57	.61	.56	.62	.59	.61	.54	.59	.59
.82	.77	.79	.79	.76	.71	.82	.78	.66	.8	.69	.8	.57	.79	.72
.85	.77	.82	.89	.83	.81	1	.83	.8	.87	.84	.83	.81	.83	.86
1	1	1	1	1	1	1.19	1	1	1	1	1	1	1	1

Table 3.12: The Power for Test Data Generated by 3PL Model with True Item Parameters

Item	$N = 500$		$N = 1000$		$N = 5000$	
	2PL	1PL	2PL	1PL	2PL	1PL
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	1	1	1	1	1	1
4	1	1	1	1	1	1
5	1	1	1	1	1	1
6	1	1	1	1	1	1
7	1	1	1	1	1	1
8	1	1	1	1	1	1
9	1	1	1	1	1	1
10	1	1	1	1	1	1
11	1	1	1	1	1	1
12	1	1	1	1	1	1
13	1	1	1	1	1	1
14	1	1	1	1	1	1
15	1	1	1	1	1	1

Table 3.13: The Power for Test Data Generated by 2PL Model with True Item Parameters

Item	$N = 500$		$N = 1000$		$N = 5000$	
	3PL	1PL	3PL	1PL	3PL	1PL
1	1	.986	1	1	1	1
2	1	.701	1	1	1	1
3	1	.562	1	.936	1	1
4	1	.372	1	.970	1	1
5	1	.042	1	.403	1	1
6	1	.079	1	.603	1	1
7	1	.006	1	.045	1	.992
8	1	.001	1	.031	1	.918
9	1	.087	1	.704	1	1
10	1	.005	1	.117	1	.986
11	1	.286	1	.670	1	1
12	1	.370	1	.788	1	1
13	1	.085	1	.556	1	1
14	.999	.538	1	.931	1	1
15	1	.930	1	1	1	1

Table 3.14: The Power for Test Data Generated by 1PL Model with True Item Parameters

Item	$N = 500$		$N = 1000$		$N = 5000$	
	3PL	2PL	3PL	2PL	3PL	2PL
1	1	.999	1	1	1	1
2	1	1	1	1	1	1
3	1	.953	1	1	1	1
4	1	.990	1	.999	1	1
5	1	.817	1	.943	1	1
6	1	.878	1	.983	1	1
7	1	.787	1	.931	1	1
8	1	.530	1	.593	1	1
9	1	.959	1	.999	1	1
10	1	.588	1	.757	1	1
11	1	.693	1	.901	1	1
12	1	.867	1	.993	1	1
13	1	.991	1	1	1	1
14	1	.898	1	.998	1	1
15	1	1	1	1	1	1

the perspective of hypothesis testing, it can be explained as that the testing of the null hypothesis (e.g., H_o here is the data fit the hypothetical 2PL or 1PL model) is being rejected almost all the times over the 1000 replications when the data are actually generated by the 3PL model under the condition of true item parameters. The rejection rate of 1 means the incorrect hypothesis is correctly rejected for each replication across three sample size conditions (500, 1000, and 5000), or the hypothesis tests for model-data misfit have perfect power.

Similarly, table 3.13 shows higher power for testing the hypothesis of fitting the data generated by the 2PL model with the 3PL model, and table 3.14 shows adequate power for testing the hypothesis of fitting the data generated by the 1PL model with the 3PL model regardless of the sample size provided item parameters are known constants.

As is known that power is a function of the sample size. As the sample size increases, power would also increase. This feature is apparent in table 3.13 and 3.14 by comparing the same hypothesis testing across three different sample sizes (e.g., 500, 1000, and 5000). For example in table 3.13 for testing the hypothesis that the item model for item 10 is the 1PL model using the data that are actually generated by the 2PL model, the power at sample 500 is .005, .117 when the sample size is 1000, and .986 when sample the size increases to 5000.

However, the power for each item is found not homogenously high, in particular for sample size of 500 case, when testing the hypothesis that the correct model is the 1PL model using the data generated by the 2PL model provided that the item

parameters are known constants. For example in the third column on table 3.13, item 3 through item 14 have very lower power for the sample size at 500. In fact, the power varies as the values of a parameter changes from item to item.

The results from table 3.13 and 3.14 also support that when fitting the data to models with more parameters than the number of item parameters for the data generating model (e.g., in table 3.13 fitting the data with the 3PL model using the data generated by the 2PL model, and in table 3.14 fitting the data generated by the 1PL model with the 3PL or 2PL model), the power is generally high provided the item parameters are known constants (except item 8 and item 10 in table 3.14).

3.4 Item Misfit and Power with Item Parameter Estimates

The simulation study in this section is similar to the above on power estimates with exception that the item parameters are not known constants but data-based estimates. When the response data are generated by the 3PL model (this is a known fact for the simulation study), then fit back the response data with the 3PL, 2PL, and 1PL models, respectively, on the basis of item parameter estimates. Lower type I error rates over 1000 replications would be expected for testing the hypothesis that the data fit the 3PL model meanwhile using the 3PL model to estimate the response data, or higher rejection rates or power would be expected when testing the hypothesis with other models (the 2PL or 1PL) meanwhile estimating the data with the 2PL or 1PL model. In addition, as seen in the above section, the power would also be expected to

Table 3.15: The Power for Test Data Generated by 3PL Model with Item Parameter Estimates

Item	$N = 500$		$N = 1000$		$N = 5000$	
	2PL	1PL	2PL	1PL	2PL	1PL
1	.182	.902	.961	1	1	1
2	.199	.796	.965	1	1	1
3	.091	.038	.935	.655	1	1
4	.563	.913	.997	1	1	1
5	.227	.218	.970	.944	1	1
6	.213	.307	.968	.963	1	1
7	.091	.136	.939	.935	1	1
8	.195	.098	.963	.826	1	1
9	.126	.661	.953	.998	1	1
10	.273	.725	.980	.999	1	1
11	.099	.062	.945	.686	1	1
12	.110	.045	.941	.688	1	1
13	.072	.644	.938	1	1	1
14	.108	.170	.943	.846	1	1
15	.656	.418	.998	.985	1	1

increase as sample size increases. Table 3.15 through table 3.17 show the power on the basis of item parameter estimates under three different data generating conditions.

One apparent characteristic in the three tables (table 3.15 through table 3.17) is that the power increases as the sample size increases. For example, when the sample size increases to 5000, the power reaches 1 at nominal level for testing the hypothesis of the 2PL or 1PL model using the data generated by the 3PL model (table 3.15), or for testing the hypothesis of the 1PL model using the data generated by the 2PL model (table 3.16). Another expected feature is that the power is generally greater when testing the hypothesis of the 2PL model (i.e., H_o: the correct model is 2PL) than the one when testing the hypothesis of the 1PL (i.e., H_o: the correct model is 1PL) given the same sample size (column 1 versus column 2 for the sample size of 500; column 3 versus column 4 for the sample size of 1000). For the sample size of

Table 3.16: The Power for Test Data Generated by 2PL Model with Item Parameter Estimates

Item	$N = 500$		$N = 1000$		$N = 5000$	
	3PL	1PL	3PL	1PL	3PL	1PL
1	.007	.494	.013	.960	.116	1
2	.059	.094	.001	.672	.117	1
3	.009	.249	.005	.709	.094	1
4	.057	.011	.000	.282	.123	1
5	.016	.002	.002	.104	.071	1
6	.016	.017	.001	.252	.074	1
7	.020	.003	.013	.004	.091	1
8	.010	.000	.003	.009	.100	1
9	.011	.050	.002	.527	.054	1
10	.017	.012	.002	.039	.128	1
11	.018	.186	.008	.571	.109	1
12	.012	.116	.006	.424	.109	1
13	.034	.008	.024	.178	.097	1
14	.007	562	.004	.950	.098	1
15	.042	.371	.001	.979	.081	1

Table 3.17: The Power for Test Data Generated by 1PL Model with Item Parameter Estimates

Item	$N = 500$		$N = 1000$		$N = 5000$	
	3PL	2PL	3PL	2PL	3PL	2PL
1	.023	.000	.024	.000	.420	.001
2	.032	.000	.019	.000	.459	.000
3	.019	.000	.020	.000	.395	.000
4	.043	.000	.025	.000	.476	.000
5	.029	.000	.023	.000	.387	.000
6	.013	.000	.014	.000	.404	.000
7	.030	.005	.032	.003	.355	.000
8	.022	.000	.023	.000	.409	.000
9	.017	.001	.020	.000	.377	.000
10	.033	.000	.022	.001	.462	.000
11	.022	.000	.029	.003	.345	.000
12	.024	.000	.013	.000	.416	.000
13	.032	.004	.027	.001	.362	.001
14	.015	.000	.016	.000	.362	.000
15	.028	.000	.024	.000	.400	.000

500, there is not enough power for testing the hypothesis of the 2PL and the 1PL using the data generated by the 3PL model except a small number of items (e.g., in testing hypothesis of the 1PL model, item 1, item 2, item 4, and item 10 seem to have adequate power that is greater or close to .80). When the sample size increases to 1000, testing both hypothesis (i.e., H_o: the correct model is the 2PL model or H_o: the correct model is the 1PL model) have power reached about .90 or greater except item 3, item 11, and item 12 when testing the hypothesis that the correct model is the 1PL model.

In table 3.16, the power for testing the hypothesis that the correct model is the 1PL model using the data generated by the 2PL model is less than .5 when sample size is 500, and there are 8 items (item 4 through item 8, item 10, item 12 and item 13) having power less .5 for testing the same hypothesis even when the sample size increases to 1000. In general, there is not enough power for testing the hypothesis of the 1PL model using the data generated by the 2PL model when item parameters are data-based estimates, in particular for the condition in which a parameters in the 2PL model are close to 1.

As is expected, the power is low for testing the hypothesis of the correct model with more item parameters than the number of item parameters for the data generating model. For example in table 3.16, the power would be low when the hypothesis is H_o: the correct model is the 3PL as compared to the 2PL data generating model no matter what the sample size is. That is to say, the item fit analysis does not have enough power to reject the test for the hypothesis that the data generated the 2PL

model fit with the 3PL model most times over 1000 replications. Similarly, table 3.17 demonstrates the item fit analysis results does have enough power to reject the hypothesis that the correct model is the 3PL or 2PL model using the data generated by the 1PL model when item parameters are data-based estimates.

3.5 True Asymptotic Distribution Versus the Approximation

The plot of the true asymptotic probabilities based on the full covariance matrix versus the approximation of the probabilities based on the observed covariance matrix among pseudocounts is very intuitive on how well the approximation works across sample sizes, with plots along the reference line $y = x$ indicating the small difference between the true and approximated values. The plots over different sample sizes may provide practical recommendations as to how large the sample size is required for an adequate approximation. For example, the following three figures (figure 3.1 through figure 3.3) are the plots of the true asymptotic probabilities and the approximation of the true asymptotic probabilities for item 1, item 3, item 5, and item 7 in the 15-item test over 1000 replications across three different sample sizes (500, 1000, and 5000). Similarly, the plots for other items can be displayed over 1000 replications, but are omitted here since the results on the plots are very close to these items.

As it can be seen from the three figures (figure 3.1 through figure 3.3), the plots spread wide along the middle of the reference line for the sample size of 500, getting narrower for the sample size of 1000, and becoming almost a straight line when the

Figure 3.1: True Asymptotic Probabilities Versus Approximation ($N = 500$)

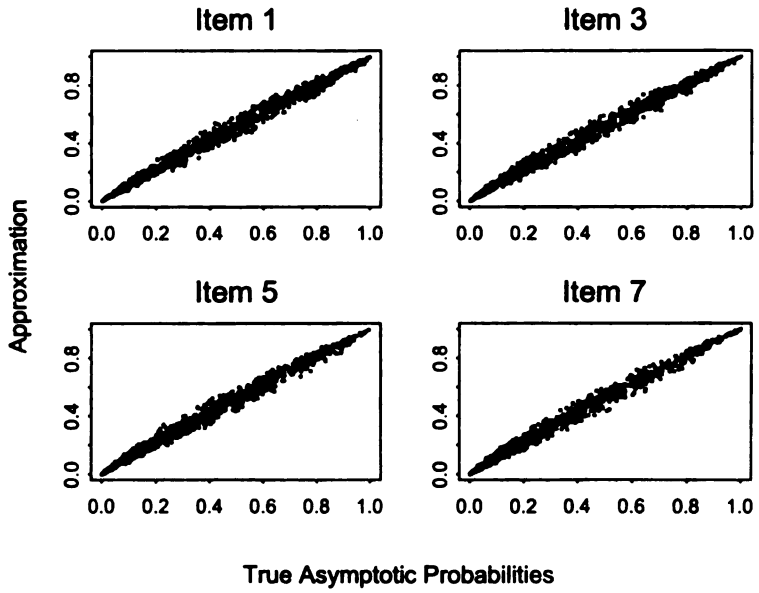


Figure 3.2: True Asymptotic Probabilities Versus Approximation ($N = 1000$)

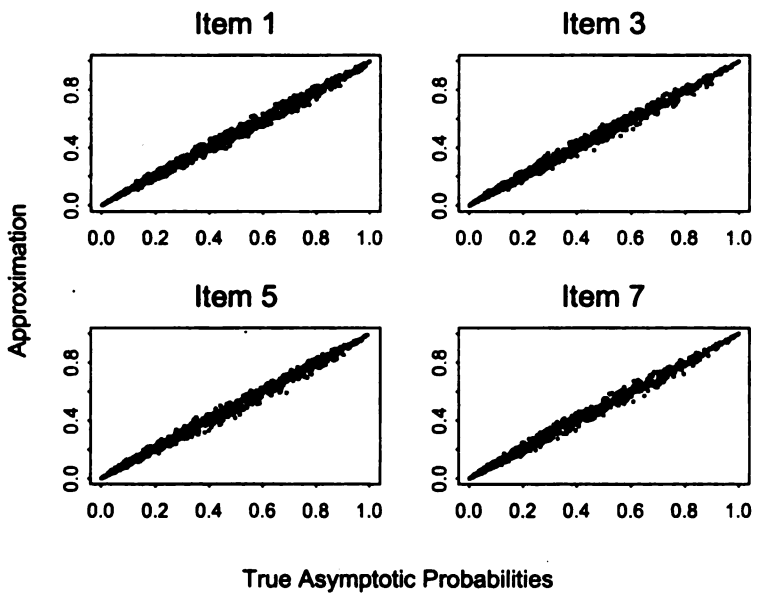
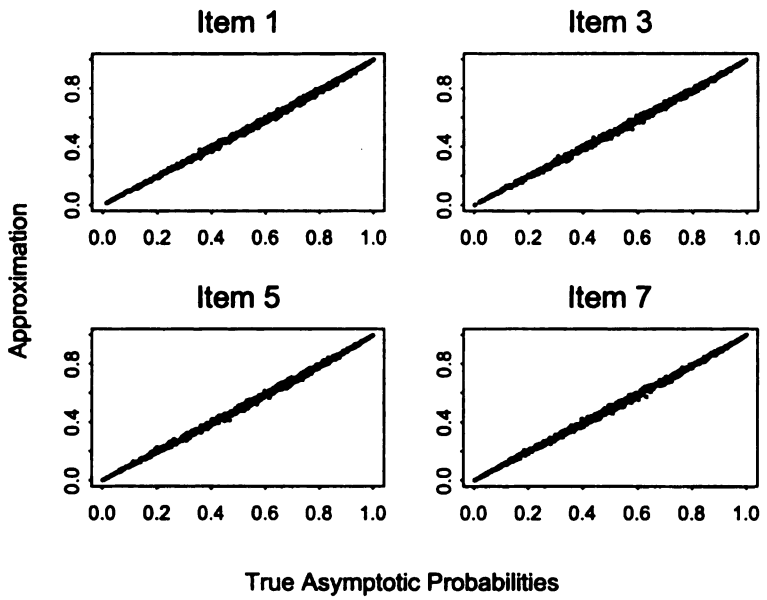


Figure 3.3: True Asymptotic Probabilities Versus Approximation ($N = 5000$)



sample size increases to 5000. Obviously, the approximation based on the observed covariance matrix of interrelations among pseudocounts works well for sample size 1000 and 5000 cases. The results on the plots are a bit dispersed for 500 examinees. In the case of a short test with small the sample size (e.g., 500), it is advised to use the true asymptotic probability instead of the approximated one.

3.6 Sensitivity Analysis

3.6.1 Non-normal Proficiency Populations

Psychometrician will be interested in finding out the applicability of one method developed in certain contexts to various other psychological and educational testing contexts. For example, in the the above studies, the group of examinees is assumed coming from a standard normal population (i.e., $N(0, 1)$), which is typically seen in

simulation studies. How does the method work with a non-normal population? This is an interesting practical issue of many tests, in which examinees do not have the exact standard normal distribution. This study is to examine the effects of the ability population distribution on the asymptotic method developed in Chapter 2.

To investigate the potential effects of the underlying ability distribution, the population is chosen as four-parameter Beta distribution ranging from -4 to 4. One reason for choosing the four-parameter Beta distribution as compared to the standard normal distribution is that it is relatively convenient to manipulate the shape and range of the distribution. The following section will briefly introduce the expectation, variance, probability density function of the distribution. The type I error rates will be examined for the above 15-item test but with a non-normal population—four parameter Beta distribution.

Four-parameter Beta distribution, denoted as $B(\alpha, \beta, L, U)$, is determined by two shape parameters (α, β) and two range parameters (lower limit L and upper limit U of the distribution). Let x be a random variable from $B(\alpha, \beta, L, U)$, i.e., $x \sim B(\alpha, \beta, L, U), L < x < U$. Then the density is given by

$$f(x) = \frac{1}{(U-L)^{\alpha+\beta-1} Beta(\alpha, \beta)} (x-L)^{\alpha-1} (U-x)^{\beta-1},$$

where $Beta(\alpha, \beta)$ is the Beta function defined for $\alpha > 0, \beta > 0$ by

$$Beta(\alpha, \beta) = \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du.$$

The expectation and variance of x can be expressed as

$$\begin{aligned} Ex &= \frac{U\alpha + L\beta}{\alpha + \beta} \\ Var(x) &= \frac{(U-L)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

If $\alpha = \beta$, then x has a symmetric distribution within its lower and upper limits. For $\alpha > \beta > 0$, x is a positively skewed distribution; for $\beta > \alpha > 0$, x is a negatively skewed distribution. For $L = 0, U = 1$, the four-parameter Beta distribution reduces to the regular Beta distribution that is often presented in basic statistics text books. In particular for $\alpha = \beta = 1$ and $L = 0, U = 1$, x degenerates as a uniform distribution within 0 and 1.

Figure 3.4 is to compare four-parameter Beta distribution with the standard normal distribution. One can find that B(4,4,-4,4) and the standard normal are symmetric but obliviously have different probability distributions. The shoulder of B(4,4,-4,4) is more wide and short than that of $N(0,1)$. Also as it is known, the range of standard normal distribution is not only restricted from -4 and 4. One can see in the figure that $B(2,4,-4,4)$ is positively skewed distribution and $B(4,2,-4,4)$ negatively skewed distribution. In this study, assume the examinees coming from $B(4,4,-4,4)$ as compared with $N(0,1)$ to see if the ability distribution has substantial effects on the results of the item fit analysis.

Table 3.18 shows that even if the underlying ability distribution is not normal, the item fit test still has low type I error rates, which is also conservative as seen in the case of the standard normal population. Again, the Bayesian procedure with

Figure 3.4: Beta Distribution versus Standard Normal Distribution

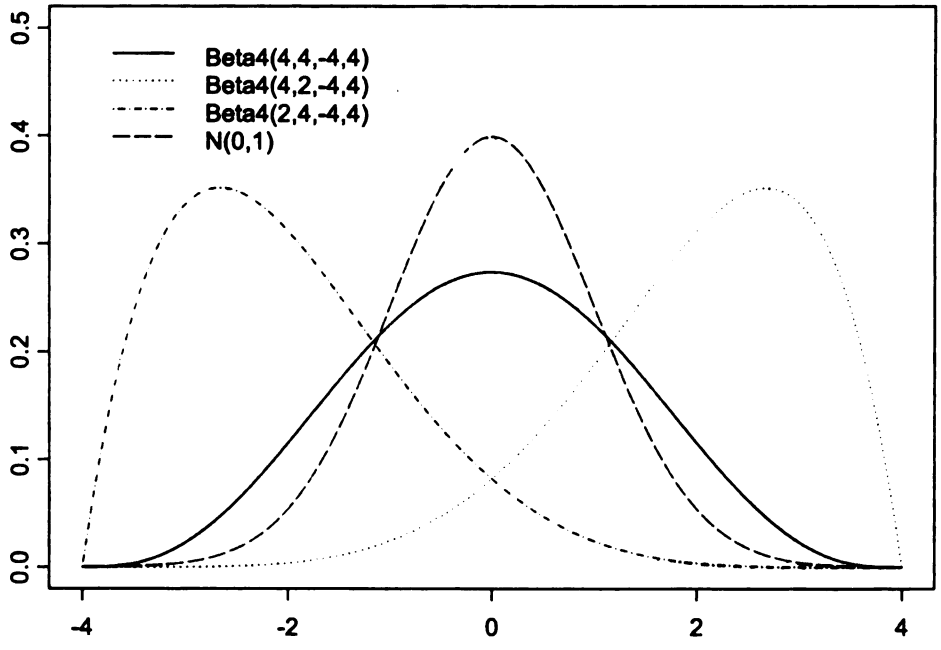


Table 3.18: Type I Error Rates for Non-normal Ability Population and Data-Based Item Parameter Estimates

Item	N=500	N=1000	N=5000
1	.000	.000	.000
2	.004	.000	.002
3	.004	.001	.040
4	.002	.000	.001
5	.000	.000	.000
6	.001	.001	.002
7	.003	.000	.000
8	.000	.000	.001
9	.000	.001	.001
10	.000	.000	.001
11	.001	.001	.000
12	.003	.001	.002
13	.010	.005	.004
14	.000	.001	.002
15	.003	.000	.000

MML is used to calibrate all item parameters with default item prior distributions when calibrating the item parameters with the 3PL model using the data generated by the 3PL model. It can be seen from table 3.18 that when the underlying ability distribution is different from the standard normal distribution, the method still provides low type I error rates, which in some sense are also viewed too conservative. The results show that the method is robust regarding the underlying ability distribution, although the item parameter estimates contains large errors in the case of the non-normal ability population. Further evidences can easily found from the RMSE for each item parameter estimate in table 3.19. The RMSE for each item in the test on three different sample sizes ($N = 500, N = 1000$, and $N = 5000$) over 1000 replications are generally larger than those RMSE in the case of the standard

Table 3.19: RMSE for Non-normal Ability Population

Item	RMSE $N = 500$			RMSE $N = 1000$			RMSE $N = 5000$		
	a	b	c	a	b	c	a	b	c
1	.3	.374	.032	.276	.365	.03	.175	.377	.025
2	.692	.357	.017	.699	.374	.009	.488	.375	.003
3	.487	.156	.066	.47	.155	.047	.313	.207	.021
4	.4	.423	.024	.359	.428	.018	.319	.416	.009
5	.553	.241	.024	.559	.237	.016	.393	.259	.009
6	.565	.208	.024	.6	.193	.016	.437	.224	.01
7	.873	.116	.029	.493	.174	.035	.384	.089	.034
8	.476	.259	.025	.473	.259	.017	.337	.278	.009
9	.618	.101	.03	.682	.073	.019	.504	.116	.012
10	.343	.413	.025	.323	.417	.022	.218	.403	.014
11	.326	.211	.03	.325	.257	.039	.247	.147	.04
12	.448	.236	.044	.426	.242	.028	.272	.28	.009
13	.667	.174	.06	.751	.234	.052	.565	.134	.024
14	.345	.124	.054	.34	.125	.048	.228	.068	.026
15	.512	.317	.023	.61	.311	.017	.508	.31	.008

normal ability distribution. The RMSE for each item parameter from the sample size $N = 500$ are indicated in the first three columns in table 3.19 corresponding to discriminating, difficulty, and asymptote item parameters, respectively. Similarly, the RMSE in the second three columns in table 3.19 for the sample size 1000, and the last three column are the RMSE for sample size 5000. One can see from the study that the effects of the ability distribution on the results of the item fit analysis are confounded with the item parameter estimation. The conservative type I error rates show that the population distribution itself should not be a factor on the results on item fit analysis, but that it can severely influence the item parameter estimates, as are represented by the large RMSE in table 3.19.

3.6.2 The Number of Quadrature Points and Item Fit

The item fit measure Q^*_{DM} or the corresponding asymptotic distribution relies on the discrete underlying ability distribution, (i.e., $p(\theta = \theta_q) = w_q$ for $q = 1, 2, \cdots, Q$), which is used to approximate a continuous distribution $N(0, 1)$. Here Q represents the number of quadrature points. How the item fit diagnostic procedure depends on the number of quadrature points Q is an important practical issue regarding the stability of the method. As is known, for a large number of quadrature points, the approximation for the distribution of the discrete proficiency gets closer to the continuous proficiency distribution. For the previous simulation studies, the number of quadrature points Q was chosen as 41 ranging within -4 and 4. To compare the stability of the results between different numbers of quadrature points, 21 and 81 quadrature points are selected within the range of -4 and 4, with similar results for the same data indicating the method is stable regarding the number of quadrature points. In this simulation study, a test of 30 items are simulated and administrated to a sample of 1000 examinees from a standard normal population. The dichotomous response data are simulated using the 3PL model. For a given data set and a stable method in which the number of quadrature points does not have substantial effects on the item fit analysis, each item fit statistic and its corresponding asymptotic probability would not expect to have a big difference as the number of quadrature points changes from 21, 41, to 81. Similarly, the type I errors rates at nominal level over 1000 replications would also not be expected to differentiate as the number of the quadrature points vary. Table 3.20 shows the true item parameters in the first three

columns of the table and the RMSE in the second three columns and type I error rates in the last three columns when $Q = 41, Q = 21$, and $Q = 81$, respectively. The item parameter estimates are MML estimates using the 3PL model in BILOG-MG3. The true item parameters in this study have a wide variety values, which intends to simulate more general practical contexts for the test items. The discriminating power parameter ranges from the smallest of .139 to the highest of 2.67; the difficulty parameters are ranging from -1.821 to 2.233; most of the asymptote parameters are around .2 with the highest of .29.

Figure 3.5 through figure 3.8 show the results on the three different numbers of quadrature points (e.g., Q=21,41, and 81). It can be seen from these figures that the plots of both the item fit statistics (i.e., Q^*_{DM}) and the corresponding asymptotic probabilities on the four items (e.g., Item 1, Item 3, Item 5, and Item 7) over 1000 replications are closely around the reference lines $y = x$, indicating these values are very close to each other no matter what the number of quadrature points is. However, with careful examination, one can find that some places are a bit messy on the plots of $Q = 21$ versus $Q = 41$, implying that some large differences occur. Similar results are also obtained from other items in the same test but not listed and plotted here. These results show the item fit analysis based upon psedocounts approach developed in Chapter 2 is not overly sensitive to the number of quadrature points, indicating a stable and robust results achieved. From these nearly interchangeable results on item fit statistics and the corresponding asymptotic probabilities, one can conclude that the number of quadrature points, practically, is not a factor that affect the results

Table 3.20: Type I Error Rates for Three Numbers of Quadrature Point

Item	True			RMSE			Type I Errors		
	a	*b*	*c*	*a*	*b*	*c*	41	21	81
1	1.899	-.054	.24	.254	.077	.036	.000	.000	.000
2	1.411	1.107	.243	.235	.097	.026	.000	.000	.000
3	2.656	-1.326	.255	.520	.136	.077	.002	.002	.002
4	2.159	1.083	.057	.279	.062	.011	.001	.001	.001
5	1.545	.735	.048	.178	.063	.018	.000	.000	.000
6	2.605	.619	.273	.415	.064	.026	.000	.000	.000
7	.771	.416	.016	.159	.162	.071	.007	.006	.007
8	2.474	1.18	.085	.362	.065	.011	.000	.000	.000
9	.941	.096	.022	.159	.137	.067	.004	.004	.004
10	2.423	.708	.246	.383	.065	.023	.000	.000	.000
11	.653	.35	.129	.119	.170	.056	.000	.000	.000
12	1.543	-.088	.226	.195	.084	.039	.003	.003	.003
13	1.832	.559	.239	.259	.072	.027	.000	.000	.000
14	1.959	.536	.096	.226	.056	.018	.001	.001	.001
15	2.587	-1.821	.096	.506	.109	.088	.002	.002	.002
16	.241	.135	.115	.166	.872	.170	.001	.003	.001
17	2.117	.838	.146	.286	.061	.018	.001	.001	.001
18	1.045	-.19	.037	.158	.128	.068	.012	.012	.012
19	.139	.211	.286	.113	.459	.048	.023	.023	.023
20	.474	1.879	.164	.178	.219	.045	.000	.001	.000
21	1.39	1.522	.222	.269	.121	.022	.000	.000	.000
22	1.972	-.963	.028	.316	.097	.082	.025	.023	.025
23	1.635	.558	.233	.229	.076	.028	.001	.001	.001
24	.381	.877	.29	.126	.312	.058	.003	.003	.003
25	.795	-.329	.197	.108	.138	.048	.002	.002	.002
26	.174	2.233	.078	.293	.439	.193	.009	.009	.009
27	1.69	2.211	.014	.297	.166	.006	.000	.000	.000
28	2.195	1.435	.066	.340	.078	.010	.002	.002	.002
29	1.268	-.331	.077	.151	.095	.050	.008	.007	.008
30	2.675	-.139	.094	.331	.052	.023	.002	.002	.002

on item fit analysis. Further evidence for this conclusion can be seen from the type I error rates at nominal level over 1000 replications in table 3.20. The largest difference of type I error rates at nominal level is .002 on item 22 (i.e., the type I error rates is .025 for Q=41 and Q=81, and .023 for Q=21), which can be attributable to the random errors of the sample data. One can use the results in this simulation study to reduce the computational complexity for a large data set since computing Q^*_{DM} based on Q=81 takes less time than the computation when Q=41. However, it is not advised to using a smaller number of quadrature points (e.g., $Q = 21$) in applications since, in a small number of cases, large disturbances occur when Q getting smaller. When Q greater or equal to 41, Figure 3.7 and Figure 3.8 show stable results on both Q^*_{DM} and its asymptotic probabilities. Therefore, $Q = 41$ is generally recommended for computing item fit in applications.

3.7 Computing Time and Programs

Several C++ programs have been implemented for the simulation studies. Three parts of C++ programs are coded for simulating response data, computing the item fit measure Q^*_{DM} for each item in a test, and evaluating the asymptotic probabilities through Davies routine (1980). The computing time, of course, depends on both the sample size and the test length. Longer tests or large sample of examinees take more time for computing item fit measure statistics. The computing time also depends on the computer equipment. The computer that is used for this simulation study is equipped with Pentium IV processor of CPU 2.39 GHZ speed and 512 MB RAM.

Figure 3.5: Item Fit Statistics Q^*_{DM} and Number of Quadrature Points

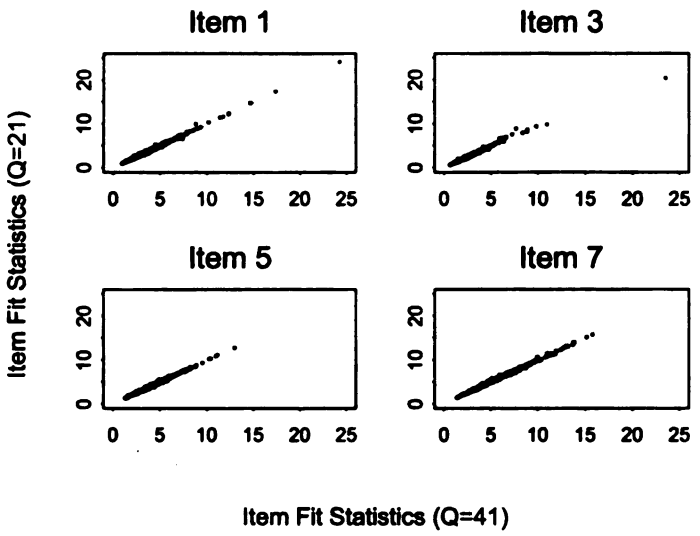


Figure 3.6: Asymptotic Probabilities and Number of Quadrature Points

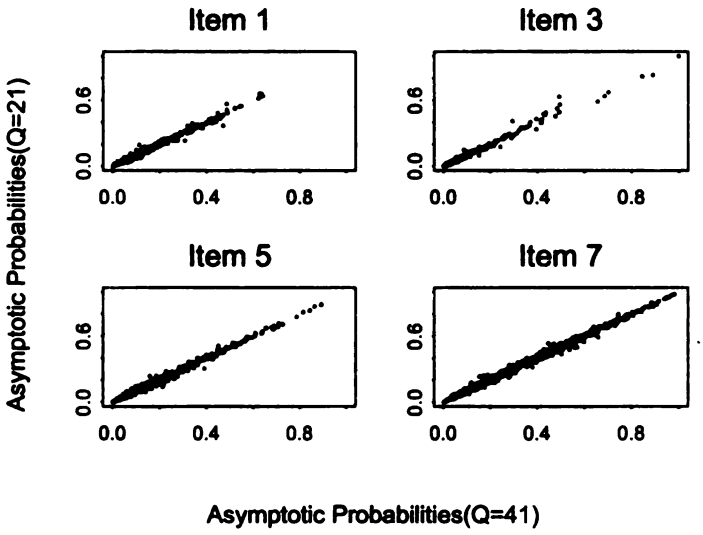


Figure 3.7: Item Fit Statistics Q^*_{DM} and Number of Quadrature Points

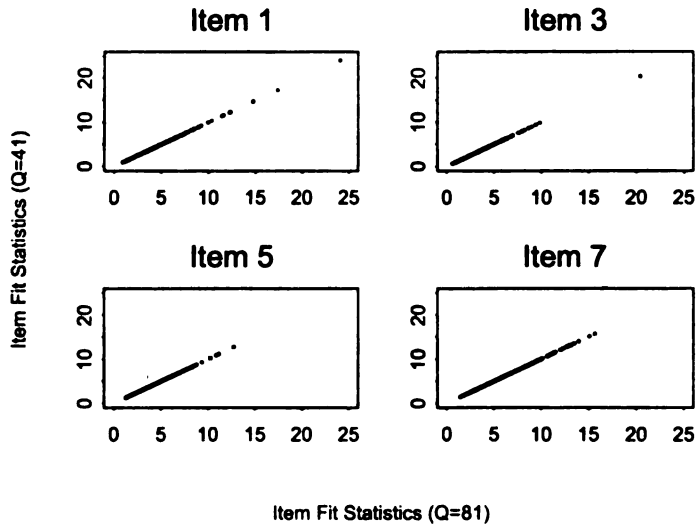
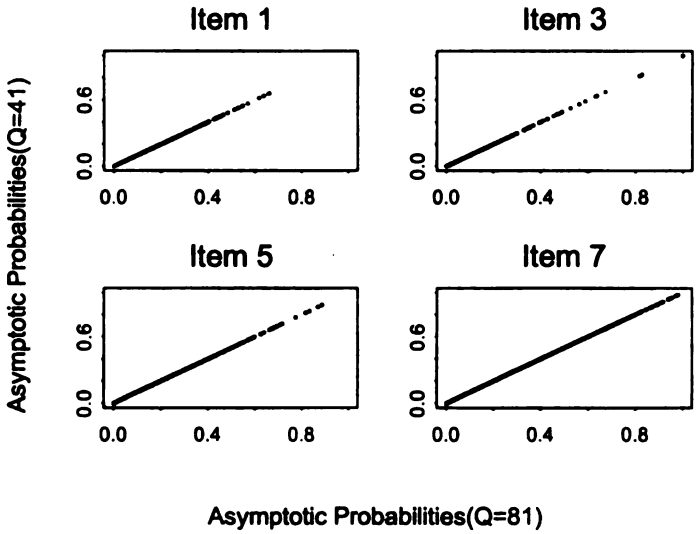


Figure 3.8: Asymptotic Probabilities and Number of Quadrature Points



The time to compute the item fit statistics (Q^*_{DM}) for each item in the test of 15 items administrated to 500 examinees takes a quarter of one minute; the time for the same test administered to 1000 examinees takes around one third of a minute; and the time for 5000 examinees takes about one and half minutes. The time to compute the test of 30 items for 1000 examinees takes around one minute. The computing time for the item fit statistics is on the basis of the number of quadrature points equal to 41. In fact, the number of quadrature points is also a factor that affects the computation time. Generally speaking, the method is robust regarding the number of quadrature points as seen in section 3.6.2. However, it takes less time for the same data set when smaller number of quadrature points is chosen. For the computation of the asymptotic probabilities, the computing time is within a second for each data set. In all the computation is efficient and applicable to most applications.

Chapter 4

Real Data Applications

One advantage of doing a simulation study on item fit analysis is that information is available about whether or not the test data fit the hypothetical IRT models. For real test data, item fit analysis is often confounded with parameter estimation (in particular for item parameter estimates) and thus make the decisions on whether or not the test data fit the hypothetical models much more complex.

4.1 Assumptions

Before doing the real data analysis, some conditions should be assumed for the sake of reasonable interpretations on the analysis results. Several assumptions that may be involved in the item fit analysis on real data. One assumption is that the parameter estimation is accurate and reliable. That is, both item and ability parameters are correctly estimated. To satisfy this condition, the standard procedures (e.g., MML for item parameter estimation and EAP for ability estimates recommended in BILOG-MG3) in most of the IRT softwares are used to estimate the parameters for the real data in this chapter. As is known, the parameter estimation is often confounded with

model-data fit issues. Poor parameter estimates may be caused by inadequate model-data fit or some other factors, for example, insufficient sample size and test length, and dimensionality or local independence conditions. Therefore, the big assumption here for real data analysis is that when testing the hypothesis that the test data fit with a hypothetical model, the parameter estimates using this hypothetical model are assumed to have no errors. For example, if one is to test the hypothesis that the observed data fit with a 3PL model, then both the item and ability parameters are correctly estimated using this 3PL model. If the parameter estimates are incorrect, the only explanation is that the test data have not adequate fit with the hypothetical 3PL model, instead of the estimation procedure itself. Other assumptions that apply to the item response theory are also all assumed here. For example, unidimensionality and local independence are assumed for the analysis in this chapter.

4.2 Two Approaches on Item Fit Analysis for Real Data

In this section for real data applications, two data sets are from Michigan Educational Achievement Program (MEAP) anonymous 2000 Fall high school science and math tests. The MEAP science data set used for this example only consists of the dichotomous responses for 19 items and 7088 examinees; the MEAP math data set here also only contains 19 dichotomous items and 6857 examinees.

In this chapter, both science and math data will be fitted in the 3PL, 2PL, and 1PL models, respectively. The item parameters will be estimated using MML method

in BILOG-MG3. The results of item fit analysis in BILOG-MG3 will be compared with the results of item fit (Q^*_{DM}) based on pseudocounts, with 10 ability groups and 30 quadrature points in the program BILOG-MG3 for χ^2 test.

Table 4.1 and table 4.2 are the item parameter estimates for the science data (table 4.1) and math data (table 4.2) corresponding to fitting the data with the 3PL model. Since the two sample sizes are large, the item fit χ^2 tests in BILOG-MG3 show that all items have statistically significant deviations between the test data and the model predictions in both science and mathematics tests (i.e., their p-values all less than .05), which are indicated by the large value of χ^2 statistics and the low p-values in table 4.1 and 4.2. As it is known, χ^2 test is sensitive to examinee sample size. Almost any departure in the data from the item model under consideration (even if the practical significance of a departure is trivial) leads to rejection of the null hypothesis of model-data fit if sample size is sufficiently large. On the other hand, for small sample size, even large discrepancies between model-data cannot be detected due to the lower power. Hambleton and Rogers (in Educational Measurement, 3rd edition, edited by Linn, (1993, p.173), "principles and selected applications of item response theory" by Hambleton) suggest that

> "statistical tests of model fit do appear to have some value. Because they are sensitive to sample size and because they are not *uniformly powerful*, the use of any of these statistics as the sole indicator of model fit is clearly inadvisable. But two situations can be identified in which these tests may lead to relatively clear interpretations. When sample size are small and the statistics indicate model misfit, or when sample size are large and model fit is obtained, the researcher may have reasonable confidence that, in the first case, the model does misfit the data, and in the second, that the model fits the data. These possibilities make it worthwhile to employ statistical tests of fit despite the alternate possibility of

Table 4.1: MEAP 2000 Fall High School Science Test Items with the 3PL Model (N = 7088)

Item	a	b	c	Q^*_{DM}	p^*	χ^2	p
1	.416	-2.742	.000	8.387	.064	53.5	.000
2	1.455	2.047	.343	2.585	.772	20.7	.023
3	.462	-1.187	.000	7.626	.093	55.8	.000
4	.598	-.181	.018	5.776	.228	64.0	.000
5	.240	-2.112	.000	3.76	.530	22.9	.011
6	.824	-.048	.198	2.681	.752	63.1	.000
7	.752	-.173	.315	2.352	.818	39.6	.000
8	.514	-1.733	.000	3.38	.606	56.1	.000
9	.556	-.664	.500	12.097	.009	37.2	.000
10	.808	.943	.256	2.411	.806	33.1	.000
11	1.048	1.255	.317	2.462	.796	31.3	.000
12	.641	-1.368	.000	4.294	.431	90.2	.000
13	.621	-1.027	.000	1.065	.796	92.5	.000
14	.635	.330	.422	3.259	.631	29.2	.001
15	.973	.588	.091	2.331	.822	107.6	.000
16	.722	.125	.325	2.548	.779	31.0	.000
17	.936	.752	.269	2.616	.765	48.8	.000
18	.747	-1.035	.000	2.888	.709	121.1	.000
19	.465	-1.783	.000	4.350	.422	68.1	.000

equivocal results."

According to the above guideline by Hambleton and Rogers, the item fit analysis results from BILOG-MG3 might not provide useful information that can lead to "relatively clear interpretation" due to the use of large sample of examinees in both tests. Or for these two examples, it is difficult for one to evaluate whether the test data on the science and math tests fit the hypothetical 3PL model if the only information available is from the results on χ^2 tests in BILOG-MG3.

Look at the results of item fit analysis for both science and math test data on the

Table 4.2: MEAP 2000 Fall High School Mathematics ($N = 6857$)

Item	a	b	c	Q^*_{DM}	p^*	χ^2	p
1	.524	-2.638	.000	4.079	.530	43.9	.000
2	.704	.074	.122	3.387	.663	38.9	.000
3	.851	1.326	.172	4.072	.531	40.3	.000
4	.771	.340	.228	3.431	.654	27.5	.002
5	.541	-2.883	.000	4.981	.379	48.6	.000
6	.912	-.173	.098	3.019	.735	58.0	.000
7	.571	-.165	.158	3.284	.683	43.0	.000
8	1.171	.559	.104	5.589	.296	57.9	.000
9	1.390	-.657	.223	3.844	.574	78.2	.000
10	.900	-.980	.198	2.849	.768	49.9	.000
11	.582	-.254	.138	3.135	.712	26.8	.003
12	1.135	-.808	.212	3.174	.705	60.4	.000
13	1.236	-.135	.226	3.400	.660	35.7	.000
14	1.329	.529	.153	4.581	.442	32.6	.000
15	.713	-1.125	.000	4.924	.387	68.0	.000
16	.414	-.620	.000	6.464	.202	49.1	.000
17	.455	-2.189	.000	8.672	.072	70.6	.000
18	.611	-.840	.500	7.110	.151	25.5	.004
19	.104	-6.671	.000	24.252	.000	77.4	.000

basis of pseudocounts. The item fit measure (Q^*_{DM}) and its corresponding asymptotic probabilities are computed using the data-based item parameters (e.g., the standard item parameter estimation procedure MML) for the science and math tests, which are the listed in table 4.1 and 4.2, respectively. It shows that item 9 in the science test and item 19 in the math test have significant deviations between the test data and the hypothetical 3PL model (i.e., p-value less than .05). Data from other items in both science and math tests are consistent with predictions based on the hypothetical 3PL model. One can also see that when the hypothetical model is being rejected, the corresponding fit statistics Q^*_{DM} is relatively larger than other items in the two tests. According to Hambleton and Roger's guideline, the test data (except item 9 for science test and item 19 for math test) have observed adequate fit with the hypothetical 3PL in both tests for such a large sample of examinees and should lead to "relative clear interpretation".

One apparently attractive property of this example of real data applications is that the item fit analysis approach based on pseudocounts (i.e., Q^*_{DM}) is able to reveal item fit test information even for the sample size as large as 7000 in this example. If both the science and math test data are fitted with the 2PL or 1PL models, then results show that all hypothesis tests for item fit analysis (Q^*_{DM}) based on pseducounts are rejected (table 4.3, 4.4). That is, the test data in both science and math test do not have adequate fit with the 2PL or 1PL models. However, different results for testing these two hypothesis are obtained from BILOG-MG3. The χ^2 tests from BILOG-MG3 shows that the test data for three items (e.g., item 1, item 5, and item 11) in

Table 4.3: MEAP 2000 Fall High School Science Items (N = 7088)

Item	2PL						1PL				
	a	b	Q^*_{DM}	p^*	χ^2	p	b	Q^*_{DM}	p^*	χ^2	p
1	.45	-2.55	17.83	.00	33.5	.00	-2.48	12.77	.00	42.6	.00
2	.12	2.52	25.44	.00	58	.00	.73	268.27	.00	209.9	.00
3	.47	-1.1	14.57	.00	38	.00	-1.1	9.82	.02	46.5	.00
4	.59	-.23	14.86	.00	48.8	.00	-.27	28.21	.00	147.5	.00
5	.24	-2.08	14.61	.00	35	.00	-1.18	9.45	.00	45.1	.00
6	.62	-.52	19.67	.00	61.4	.00	-.64	36.61	.00	18.4	.00
7	.53	-1.02	17.34	.00	35.1	.00	-1.12	16.72	.00	94.3	.00
8	.52	-1.71	14.24	.00	27.9	.00	-1.86	12.56	.00	79.2	.00
9	.39	-2.57	17.49	.00	34.2	.00	-2.21	18.38	.00	29.6	.00
10	.41	.24	24.15	.00	115.5	.00	.22	25.75	.00	101.5	.00
11	.32	.45	35.36	.00	96.2	.00	.33	63.59	.00	111	.00
12	.68	-1.32	14.34	.00	44.1	.00	-1.73	4.24	.00	161.6	.00
13	.64	-.95	14.75	.00	41.1	.00	-1.18	3.49	.00	146.7	.00
14	.36	-1.26	17.01	.00	47.7	.00	-1.02	27.38	.00	35.4	.00
15	.74	.41	18.19	.00	125.1	.00	.57	8.98	.00	369	.00
16	.46	-.85	19.4	.00	69.2	.00	-.84	14.01	.00	67.8	.00
17	.46	.04	25.73	.00	105	.00	.04	23.8	.00	122.1	.00
18	.76	-1.02	14.54	.00	78.6	.00	-1.43	67.39	.00	225.3	.00
19	.5	-1.7	14.18	.00	35.5	.00	-1.77	1.71	.01	54.5	.00

math test have reasonable fit to the 2PL model and that three items (i.e., item 1, item 11, and item 17) also in math test have reasonable fit to the 1PL model (table 4.4). Interestingly, note that in the math test the same three items (e.g., item 1, item 5, and item 11) shows reasonable fit with the 2PL model but inadequate fit with the 3PL model, which might be hard to make sense. Similarly, it is also difficult to consider a situation that the data from the same three items (item 1, item 11, and item 17) in the math test have reasonable fit with the 1PL model but fail to support the fit with the 3PL models. These results seem conflict with the general principles that the more parameters in the model the better fit may be achieved merely from the model-data fit perspective.

Table 4.4: MEAP 2000 Fall High School Mathematics Items ($N = 6857$)

Item	2PL						1PL				
	a	b	Q^*_{DM}	p^*	χ^2	p	b	Q^*_{DM}	p^*	χ^2	p
1	.56	-2.5	39.35	.00	1.2	.34	-2.39	19.71	.00	12	.22
2	.59	-.23	4.39	.00	17.4	.04	-.23	21.74	.00	41.5	.00
3	.41	1.08	56.91	.00	66.6	.00	.82	111.33	.00	83.9	.00
4	.52	-.28	48.21	.00	26.6	.00	-.25	4.26	.00	28.1	.00
5	.58	-2.74	39.7	.00	13.4	.15	-2.67	2.5	.00	17.8	.04
6	.79	-.38	4.91	.00	67.2	.00	-.45	62.3	.00	163.1	.00
7	.48	-.62	39.7	.00	18.7	.03	-.53	4.03	.00	25	.00
8	.81	.36	56.25	.00	108.9	.00	.45	72.82	.00	207.4	.00
9	1.1	-1.01	43.97	.00	47.7	.00	-1.43	152.57	.00	337.2	.00
10	.8	-1.32	4.16	.00	36.5	.00	-1.6	55.26	.00	13.8	.00
11	.5	-.64	41.78	.00	13	.16	-.57	32.86	.00	11.4	.25
12	.95	-1.16	43.77	.00	4.2	.00	-1.53	103.8	.00	241.8	.00
13	.85	-.59	51.91	.00	76	.00	-.73	77.2	.00	212.1	.00
14	.77	.24	68.08	.00	142.4	.00	.29	71.94	.00	206.7	.00
15	.74	-1.08	39.04	.00	39.3	.00	-1.24	37.67	.00	96.3	.00
16	.44	-.62	41.16	.00	43.8	.00	-.5	57.35	.00	32.6	.00
17	.48	-2.11	41.29	.00	23.5	.01	-1.79	35.81	.00	14.3	.11
18	.47	-2.28	48.98	.00	23.6	.01	-1.92	42.09	.00	2.6	.01
19	.1	-7.2	51.48	.00	67.4	.00	-1.39	514.21	.00	42.4	.00

4.3 Graphic Approach

One more interesting question is what other evidence one can have to further support the assessment decisions on the apparently different results from the above two approaches (i.e., χ^2 test and Q^*_{DM}) on item fit analysis. One alternative approach—graphic approach—might provide some intuitive sense to help assess on whether or not the test data from MEAP science and math tests fit the hypothetical IRT models.

Figure 4.1 through figure 4.5 are the plots of the hypothetical 3PL model item response function (denoted as solid curve in the graph) with the observed empirical item response curve (denoted as dot in the graph) for the 19 items in the science test. One can see from these plots that most of the items do have reasonable fit with the

3PL model, assuming the estimation is correct and other assumptions for IRT (e.g., local independence and unidimensionality) are satisfied. Item 9 is diagnosed to have significant deviation between the data and the 3PL model, which can be seen in the first plot on figure 11 with large discrepancy (i.e., more .5 deviation) between the hypothetical IRF and the emprical IRF at the lower end of ability scale. In fact, it can be seen that there are other items (item 1, item 3, item 5, item 8, and item 19) that also show large discrepancies at the lower end of the ability scale but result in reasonable fit. One possible explanation to this finding is that there may have large errors for the ability estimates, which lead misclassifications for examinee groups. The reason for the possible large errors for ability estimation, in particular for the ability estimates at the two ends on the ability scale, may be attributable to the small number of test items in the science test (i.e., a 19-item test can consider to be a short test). That is the part of the reasons why BILOG-MG recommends using χ^2 test for a test with more than 20 items. Combined with the plots diagnose and results on item fit analysis, one can conclude that the Q^*_{DM} test provides helpful information on assessing model-data fit. Moreover, the Q^*_{DM} test for item fit can apply to short tests and large sample of examinees, which broaden the settings for item fit analysis.

Figure 4.1: Empirical versus Hypothetical Item Response Functions for MEAP 2000 High School Science Items (1-4)

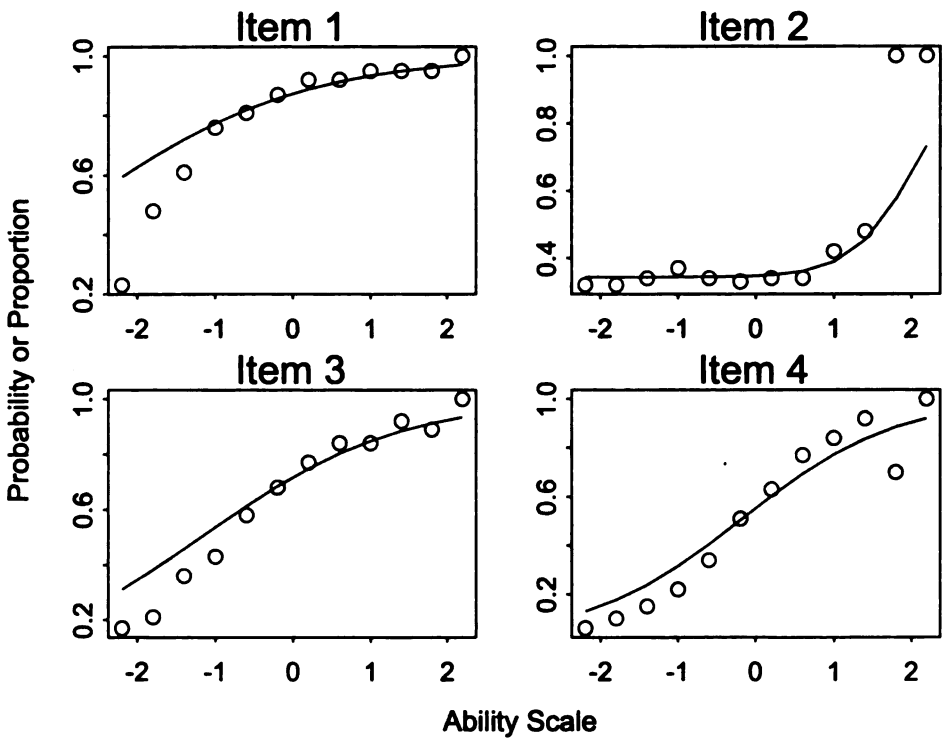


Figure 4.2: Empirical versus Hypothetical Item Response Functions for MEAP 2000 High School Science Items (5-8)

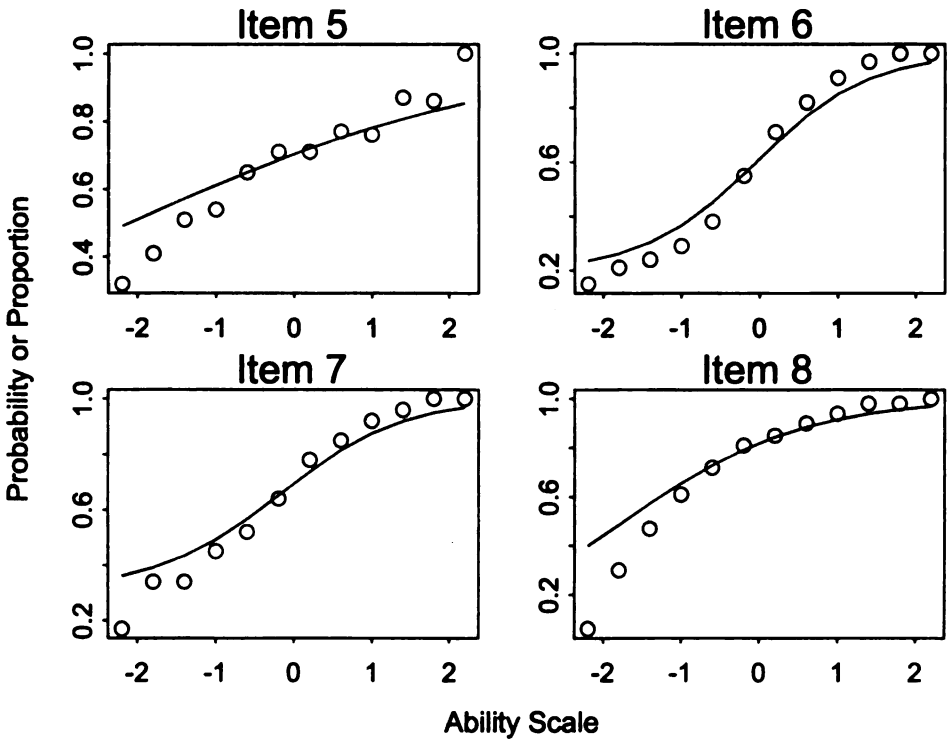


Figure 4.3: Empirical versus Hypothetical Item Response Functions for MEAP 2000 High School Science Items(9-12)

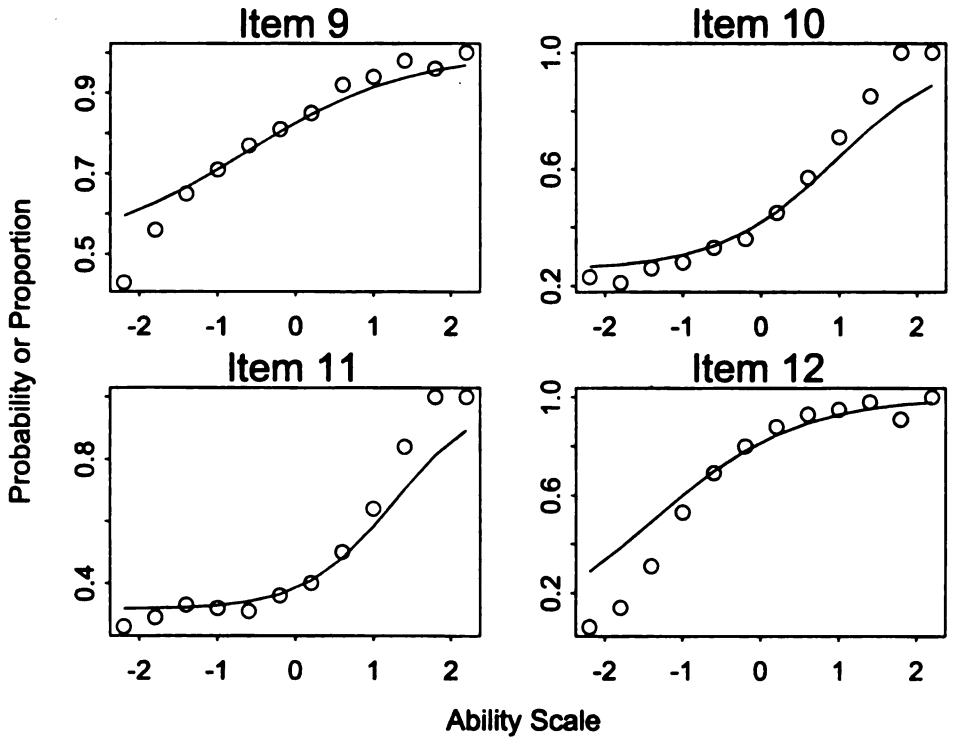


Figure 4.4: Empirical versus Hypothetical Item Response Functions for MEAP 2000 High School Science Items(13-16)

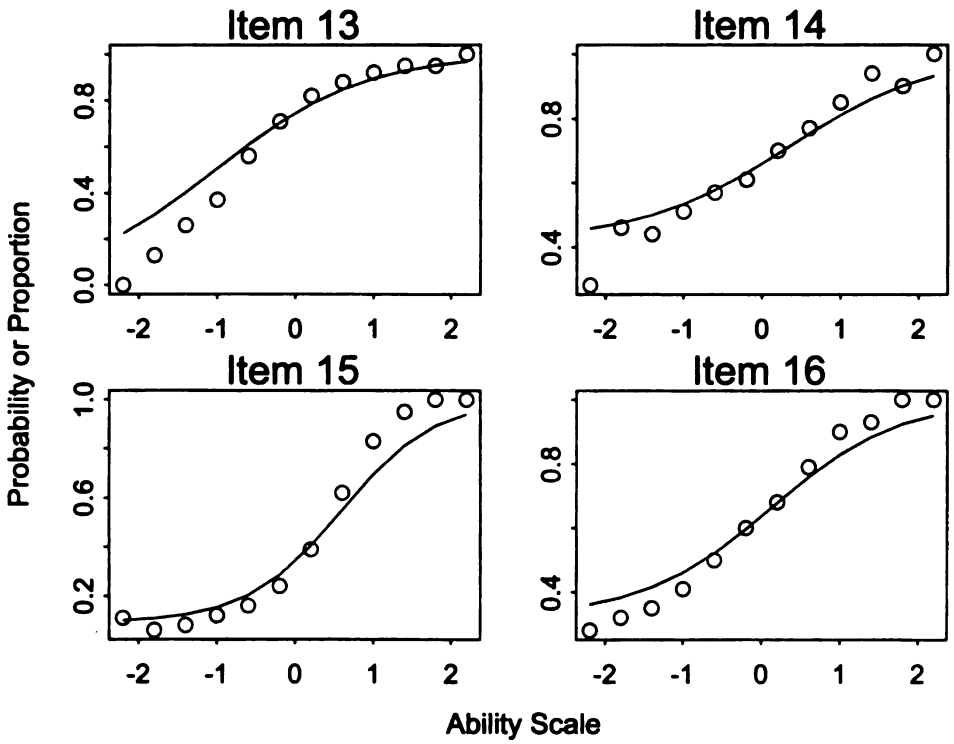
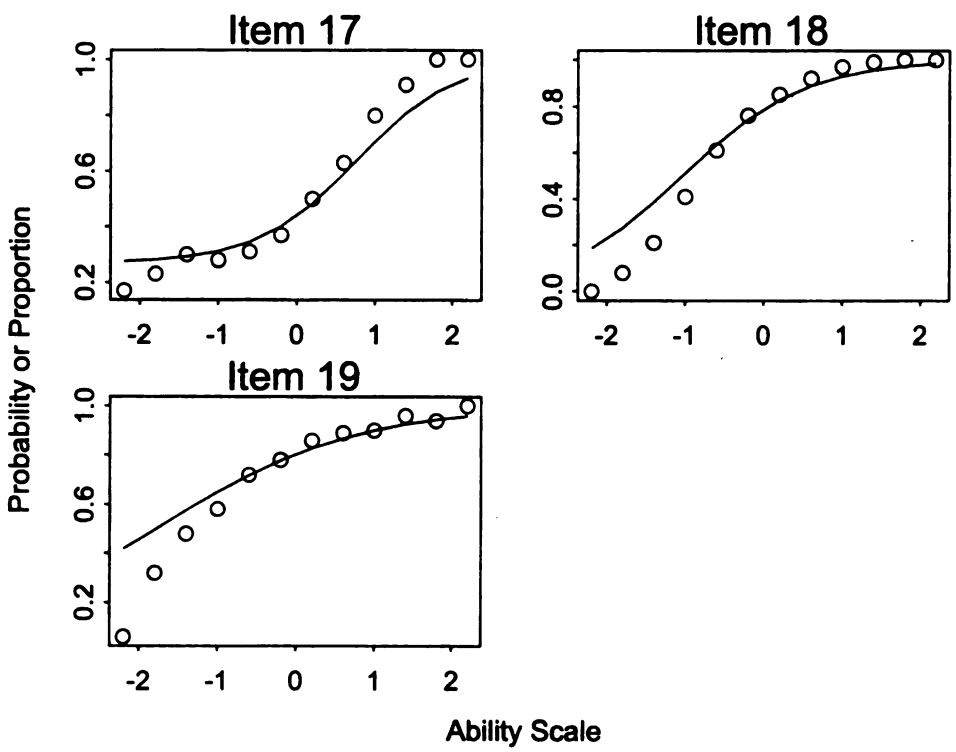


Figure 4.5: Empirical versus Hypothetical Item Response Functions for MEAP 2000 High School Science Items(17-19)



Chapter 5

Concluding Remarks and Future Research Directions

The simulation studies in Chapter 3 demonstrate that the approach to detect item fit or misfit is reliable and promising. The approach achieves the expected computational efficiency by approximating the true asymptotic probabilities based on the observed covariance matrix of interrelations among pseudocounts (e.g., Figure 1), thus making the approach applicable to most operational research. The approximation not only brings computational simplification, but also produces accurate results on assessing item fit from the oracle analysis in Chapter 3. When other sources of errors are controlled, for example in the condition if item parameters are known, the item fit test statistic Q^*_{DM}, the coefficients of the asymptotic distribution (table 3.5 through table 3.11), the asymptotic probabilities (Figure 1 to 3), type I error rates (table 3.2, 3.3, 3.4), and the decisions on whether the test data fit the hypothetical models have good agreement on the basis of the approximation. However, it is a fact that the approximation based on the observed covariance matrix among pseudocounts brings additional errors for assessing item fit, and the error may be large in the situation

when the test is long and only a small number of examinees is available.

The utility of this approach is not limited to test length. For short tests, for example, a test with 10 items or less, one can directly use the true asymptotic distribution rather than its approximation to evaluate whether or not the test data fit the hypothetical model, because computing the true asymptotic distribution only needs to evaluate 1024 possible response patterns no matter how large the sample size is. However, it is advised to use a sample at least as large as 1000 to achieve better approximation. It can be seen from Figure 1 that the approximation looks a bit dispersed when sample size is 500, but is improved when the sample size increases to 1000.

This approach has strong theoretical basis, because the fundamental concept of this approach is "pseudocounts," or the posterior of ability distribution instead of ability estimates, which is believed to provide better information on assessing item fit. One direct theoretic advantage of using "pseudocounts" rather than "ability estimates" to evaluate item fit is that this approach is able to avoid additional sources of errors that are confounded with ability estimation in item fit analysis, in particular for short tests. For a short test, the ability scale might not be well defined, and thus the large errors induced by ability estimates and classification errors by grouping examinees make the results from the χ^2 item fit test questionable, as is the case for the example on Chapter 4 real data applications. But this is not a problem on the approach based on pseudocounts because the observed counts from ability estimates are not required for the analysis. The following is the summary on other advantages

and limitations.

First of all, the approach of detecting item fit has reasonable type I error rates (table 3.2, 3.3, 3.4, 3.18, 3.20). In table 3.2, 3.3, and 3.4, when the item parameters are known constants, one can see that the type I error rates ranges from 0 to .05 with most items having type I errors rates around .02, which is acceptable. However, when the item parameters are data-based estimates, almost all items have conservative type I error rates no matter what the sample size is and how good the item parameter estimates are in the analysis. The too conservative type I error rates when item parameters are estimated are resulted from the under estimates of the item fit statistics Q^*_{DM}, which also lead to under estimates of the corresponding asymptotic probabilities. In Chapter 2, it is addressed that the asymptotic distribution can be expressed as a linear combination of the independent χ^2 variables. The coefficients on the basis of item parameter estimates for the linear combination on each item are arbitrarily close to those from the true item parameters (see table 3.5 though 3.11).

One interpretation to the conservative type I error rates for the data-based item parameter estimates can be attributable to the estimation errors (e.g., errors for estimating the covariance matrix, errors for estimating the eigenvalues, and errors for estimating item parameters), which result in under estimates of the item fit statistic (Q^*_{DM}) and its asymptotic probability. Since the eigenvalues seem well estimated by table 3.5 through table 3.11, the conservative type I error rates could resulted from the under estimates of the item fit statistic due to the errors for estimating item parameters. Note that the extension of the results on item fit analysis to the context

of item parameter estimates relies on the availability of consistent estimates for item parameters. Although the RMSE for item parameter estimates when the sample size is 5000 are much smaller than those when the sample size is 500 and 1000 (see table 3.2 through 3.4), the estimates of item parameters contain a large amount of errors for each item. If the item parameters would not contain estimation errors, one could expect the similar type I error rates to those when the item parameters are known constants. It is possible that poorly recovered item parameters from observed data cause the poor item fit results in the simulation studies. Therefore, it is necessary to discern if poor item fit is resulted from that the test data really inadequately fit the item models or from the item parameters that are poorly estimated possibly due to bad estimation procedures. That is, although the item fit analysis on the situation when the item parameters are data-based estimates does not rely on ability estimates, detecting item fit or misfit based on pseudocounts requires item parameter estimates, which inevitably confounds the model-data fit issues with the estimation issues. Poorly recovered item parameters lead to questionable model-data fit analysis. It is also true that inadequate model-data fit will result in questionable item parameter estimates. Further research work is still needed to investigate the effects of item parameter estimates on the model-data fit analysis. For example, further efforts are needed to examine what cause the under estimates of the item fit statistics and how to correct the effects of item parameter estimates. One possible approach is found in Donoghue & Hombo (2003) by explicitly examining the effect of item parameter estimation and deepening the understanding of its effect on the distribution of item

fit measure.

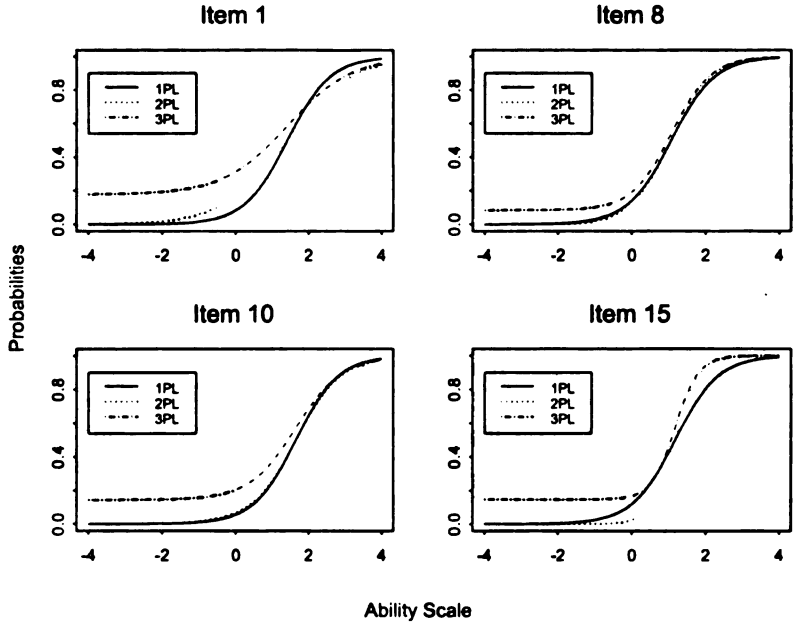
Secondly, the approach has adequate power to detect item misfit (table 3.12, 3.13, 3.14, 3.15, 3.16, 3.17) in the simulation studies. When item parameters are true values, the power estimates for many items are around .9 even when the sample size is as small as 1000 (see table 3.12 to 3.14). Item 5 through item 13 in table 3.13 and item 8 and item 10 in table 3.14 show that the power less than .9 and varies across these items as their discriminating power (a parameters) get closer to 1, which can be explained by the relations between their item response functions. In general, when item parameters are true values, the more separation of the IRF between the true model and the hypothetical model, the easier to detect item misfit, and the higher power could be expected for even small sample size (e.g., 500). For example, the 3PL model can be more likely to be separated from the 2PL or the 1PL model because of the presence of the asymptote parameters. However, the 2PL and the 1PL model can hardly be separated from each other in particular when the discriminating power parameters are close to 1 and the 2PL model nearly reduce to the 1PL model, which is also difficult to detect from test data. Therefore, to detect misfit on the 2PL or 1PL, the power should be a function of item discriminating power parameter and the power curve over a parameter can be expected to look like a "U" shaped curve with the lowest power associated with a parameters close to 1.

The item response function can provide information for diagnosing item fit testing process. In the simulation process, it is the item response function that determines the simulation of the dichotomous response data. In Chapter 2, it is also shown

how an IRF influences the pseudocounts, the sum of the posteriors over all possible response patterns for the rest items in a test, and how an IRF directly affects the theoretic expectation of pseudocounts, and eventually how an IRF influences on the item fit measure Q^*_{DM} and its corresponding asymptotic distribution.

If the two IRF are very close to each other, one can expect that the two models would fit a data set equally well or would not have reasonable fit for the data at the same time. Thus the power may be low in the situation when the two IRF are close, and large sample size may be required to detect the misfit. For example, the 2PL model and the 1PL model have the same asymptote value. When an IRF from a 2PL model has very similar curve to an IRF from a 1PL IRT model (or the a parameter for the 2PL model is close to 1), and if the data can reasonably fit the hypothetical 2PL model, the data can also be expected to fit well for the hypothetical 1PL model, and vice versa. Look at the true item parameters in table 3.1, the discriminating power parameter a's starting from item 5 to item 13 are close to 1, in particular for item 8 and item 10 with discriminating power parameters equal to 1.107 and .92, respectively. If the asymptote parameter c is disregarded, then the 2PL and the 1PL (treat all a's value as 1) IRF should have a slight difference. Therefore, although the data sets are generated from the 2PL model, the power should be low for rejecting the 1PL model (see table 3.13) due to the fact that the two IRF are too close to each other. Similarly, the power should also be low for item 8 and item 10 when fitting the data generated by the 1PL model with the 2PL model, as reported in table 3.14. Figures 5.1 shows the comparison of the 3PL, 2PL, and 1PL IRFs for item 1, item

Figure 5.1: Item Response Functions for the 3PL, 2PL, and 1PL Model (Item 1, 8, 10, 15)



8, item 10, and item 15. It is apparent that the 3PL, 2PL, and 1PL IRF for item 1 and item 15 are well separated and thus these two items have higher power even for sample size 500. On the other hand, the 2PL and 1PL curves for item 8 and item 10 are too close to separate from each other, as seen in the figure, and thus have lower power for the sample size as large as 1000. However, their IRF are well separated from the 3PL model and these two items also observe higher power for detecting misfit of the 3PL model.

In a short, for detecting item misfit, an IRF from a 3PL model can be easily separate from an IRF from other models (e.g., the 2PL and the 1PL). Therefore, this is why higher power is observed when fitting a 2PL or 1PL model using the data

generated by a 3PL model. However, when the test data are generated by a 2PL model, if a hypothesis of fitting the data with a 1PL model and the two IRF are not well separated, it is hard to expect adequate power unless a sufficiently large sample size is available.

When item parameters are data-based estimates, the power for detecting misfit (e.g., the 2PL or 1PL model) for most items is .9 or greater when data are generated using the 3PL model and the sample size is large (1000), as seen in the third and fourth column in table 3.15. The lowest power for three items (item 3, item 11, and item 12) has power around .7. However, when data are generated from the 2PL model, the test for fitting the data with the 1PL model shows very low power, which can be seen in table 3.16, in particular for item 4 through item 13, whose IRF are close to that of the 1PL model.

Next, the method is robust in terms the ability distribution, and is insensitive to the change of the number of quadrature points. Although the results on type I error rates in table 3.19 with non-normal ability population (i.e., Beta distribution in the example) show that the method is robust over the underlying ability distribution, too conservative type I error rates are observed with poorly recovered item parameters, as can be seen from their root mean square errors. Here the problem of non-normal ability population turns back to the discussions on the effects of item parameter estimates on the item fit analysis. As is true that poorly recovered set of item parameters cannot yield a correct decision on whether or not the test data fit the hypothetical item models even in simulation studies, it is also true that the results

on item fit analysis based upon the bad item parameter estimates may not support that the test data fit the hypothetical model even though the data are generated by the hypothetical model. That is, one can obtain unacceptably high type I error rates using a set of bad item parameter estimates. The point is that how bad are the item parameter estimates can be tolerated for the use of the results from item fit analysis. The study of non-normal ability population only provides a general sense of how the bad item parameter estimates can have effects on the item fit analysis in terms of the root mean square errors. In the table 3.19 on RMSE for the non-normal ability population across three different sample sizes (500, 1000, and 5000), one can see that most of the RMSE for discriminating power parameters are greater than .5, for difficulty parameters greater than .3, and for asymptote parameters greater than .03. More research work is needed to study the tolerance of the item fit on the effects of item parameter estimates.

As for the effects of the number of quadrature points on the results of item fit analysis, it can be seen in table 3.20 and from Figure 3.5 through Figure 3.8, the results based on $Q = 21$ and $Q = 41$ have slight differences. However, the results based on $Q = 41$ and $Q = 81$ show extremely good consensus. Thus, it is advised to compute item fit analysis using 41 quadrature points to have both computing accuracy and efficiency.

Finally, although the method takes an asymptotic approach, it works extremely well even for the sample size of 1000 and the test of item fit is not sensitive to the number of examinee sample size. When the test has the sample size as large as

5000, χ^2 test for item fit will tend to reject the hypothesis on most items, whereas Q^*_{DM} statistic test will still provide useful information on diagnosing item fit, as is evident in table 3.2 through table 3.4 and in the example of real data applications in Chapter 4. The high school science and math MEAP data include large sample of examinees, which makes it hard to diagnose whether or not the test data fit the hypothetical 3PL model using χ^2, as shown in table 4.1 and table 4.2. Additional evidence from the plots between the hypothetical IRF and the empirical IRF for each item in the science test in Figure 4.1 through 4.5 show that the results from Q^*_{DM} analysis provide reliable information, which agree with the results obtained from the graphic approach. One can conclude from the real data applications that the item fit Q^*_{DM} diagnosing test is able to provide more helpful information on assessing the model-data fit.

In a short, the reformulation of the Q^*_{DM} seems not correct on the conservative type I error rates when item parameters are data-based estimates. However, the reformulation does provide a convenient theoretical framework for studying item fit based on pseudocounts.

Bibliography

[1] Birch, M. W. (1964). A new proof of the Pearson-Fisher theorem. *Annals of Mathematical Statistics*, 35, 818-824.

[2] Bishop, Y. M. M., FeinBerg, S. E., and Holland, P. W. (1975). *Discrete multivariate analysis.* Cambridge, MA: The MIT Press.

[3] Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.

[4] Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, 46, 443-449.

[5] Bock, R.D., and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.

[6] Davies, R. B. (1980). Algorithm AS 155: Distribution of a linear combination of non- central chi-squared random variables. *Applied statistics*, 29, 323-333.

[7] Donoghue, J.R., and Hombo, C. M. [McClellan, C. A.] (2003b). Some asymptotic results on the distribution of an IRT measure of item fit. *Psychometrika* (conditionally accepted).

[8] Donoghue, J.R., and Hombo, C. M. [McClellan, C. A.] (2003a, April). A corrected asymptotic distribution of an IRT fit measure that accounts for the effects of item parameter estimation. *Paper presented at the Annual Meeting of the American Educational Research Association*, Chicago, IL.

[9] Donoghue, J.R., and Hombo, C. M. [McClellan, C. A.] (2001b, June). The behavior of an IRT measure of item fit in the presence of the item parameter estimation. *Paper presented at the Annual Meeting of the Psychometric Society*, Valley Forge, PA.

[10] Donoghue, J.R., and Hombo, C. M. [McClellan, C. A.] (2001a, April). The distribution of an item-fit-measure for polytomous items. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, Seattle, WA.

[11] Donoghue, J.R., and Hombo, C. M. [McClellan, C. A.] (1999, June). Some asymptotic results on the distribution of an IRT measure of item fit. *Paper presented at the Annual Meeting of the Psychometric Society*, Valley Forge, PA.

[12] Donoghue, J. R., and Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement*, 22, 33-51. Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia, PA: Society of Industrial and Applied Mathematics (SIAM).

[13] Glas, C. A. W., and Meijer, R. R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied psychological measurement*, 27(3), 217-233.

[14] Glas, C. A. W., and Suarez-Falcon, J. C. (2003). A comparison of item fit statistics for the three-parameter logistic model. *Applied psychological measurement*, 27(2), 87-106.

[15] Hambleton, R. K., and Swanminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer Academic Publishers.

[16] Hoijtink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A. J. van duijn, and T. A. B. Snijiders (Eds.), *Essay in item response theory* (pp. 109-130). New York: Springer.

[17] Hombo, C. M. [McClellan, C. A], and Donoghue, J. R. (2001, July). A power study of an IRT measure of item fit. *Paper presented at the Annual Meeting of the Psychometric Society*, King of Prussia, PA.

[18] Hombo, C. M. [McClellan, C. A], and Donoghue, J. R. (2000, July). Some properties of the distribution of an IRT measure of item fit. *Paper presented at the 2000 annual meeting of the Psychometric Society*, Vancouver, British Columbia.

[19] Hombo, C. M. [McClellan, C. A], and Donoghue, J. R. (1999, June). A simulation study of distribution of an IRT measure of item fit. *Paper presented at the Annual Meeting of the Psychometric Society*, Lawrence, KS.

[20] Hombo, C. M. [McClellan, C. A], and Donoghue, J. R., and Oranje, A. H. (2003, April). Evaluating item fit in 2002 NAPE writing data. *Paper presented at the annual meeting of the American Educational Research Association*, Chicago, IL.

[21] Hsu, Y. (2000). On the Bock-Aitkin procedure—from an EM perspective. *Psychometrika*, 65, 547-549.

[22] Johnson, N. L. and Kotz, S. (1970). *Continuous univariate distributions* —2. Boston: Houghton Mifflin.

[23] Li, D., Donoghue, J. R., and McClellan, C. A. (2005). Approximate the asymptotic distribution of an IRT measure for item fit based on observed interrelations among pseudocounts. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, Montreal, CA.

[24] Linn, R. L. (1993). *Educational Measurement*, 3rd edition. The Oryx Press.

[25] McKinley, R. L., and Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9, 49-57.

[26] Orlando, M. and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50-64.

[27] Orlando, M. and Thissen, D. (2003). Further investigation of the performance of S-χ^2: an item fit index for use with dichotomous item response theory models. *Applied psychological measurement*, 27(4), 289-298.

[28] Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.

[29] Reise, S P. (1980). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, 14, 127-137.

[30] Rogers, H. J., and Hattie, J. A. (1987). A Monte Carlo investigation of several person and item fit statistics. *Applied Psychological Measurement*, 11, 47-57.

[31] Sinharay, S. (2005). Bayesian item fit analysis for unidimensional item response theory models. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, Montreal, CA.

[32] Sinharay, S., and Johnson, M. S. (2003). Simulation studies applying posterior predictive model checking for assessing fit of common item response theory models (ETS RR-03-33). Princeton, NJ: ETS. Available from *http://www.ets.org/research/newpubs.html.*

[33] Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of- fit statistic in IRT models. *Journal of Educational Measurement*, 37, 58-75.

[34] Stone, C. A., Ankenmann, R. D., Lane, S., and Liu, M. (1993, April). Scaling QUASAR's performance assessments. *Paper presented at the Annual Meeting of the American Educational Research Association*, Atlanta, GA.

[35] Stone, C. A., and Hansen, M. A. (2000, April). Using resampling methods to evaluate the significance of a goodness-of-fit statistics in item response theory model. *Paper presented at the Annual Meeting of the American Educational Research Association*, Montreal, Canada.

[36] Stone, C. A., Mislevy, R. J., and Mazzeo, J. (1994, April). Misclassificaiton error and goodness-of-fit in IRT models. *Paper presented at the Annual Meeting of the American Educational Research Association*, New Orleans, LA.

[37] Stone, C. A., and Zhang, B. (2002). Comparing three approaches for assessing goodness- of-fit of IRT models. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, New Orldeans, LA.

[38] Yen, W. M. (1981). Using simulation to choose a latent trait model. *Applied Psychometrical Measurement*, 5, 245-262.